"POLITEHNICA" UNIVERSITY of TIMIŞOARA
Faculty of Automatics and Computers
Department of Computer & Software Engineering

Contributions to a GSM handsfree system operating in the time domain

Doctoral Thesis

Thesis supervisor: Prof. Dr.-Ing. Crişan Strugaru

Doctorand: Dipl.-Ing. Ingrid Kremmer

1998

Meiner lieben Dori († München, 03.05.97, 10:38)

Acknowledgments

I would like to express my respectful thanks to my dissertation supervisor, Prof. Dr.-Ing. Crişan Strugaru, for his competent assistance, his valuable instructions and for his patience with my problems during the whole elaboration period of this work.
I also thank the academic staff of the Department of Computers & Software Enginering of the Faculty of Automatics and Computers at the "Politehnica" University of Timişoara for their support.

I am much indebted to Mr. Horst Fenske of Siemens AG, Munich for making it possible to study for doctorate during my past activities in the Mobile Phones Department.

I would further like to express my thanks to the student trainees, workstudents and student graduates I had been supervising during the last two years. Here I have to mention Mr. Stefan Bayer and Mr. Falko Fiechtner for their support in implementing some algorithms under study and Mr. Bernhard Zwilling for processing the relevant statistical information regarding the current telecommunication situation. My special thanks go to my colleague Mr. Thorsten Ansahl for his help and suggestions concerning the editing in LaTeX and for proof-reading this thesis.

Last but not least, I am deeply grateful for the love, encouragement and mental support I always got from my family, especially from my mother who unfortunately has not got to see this dissertation finished.

Munich, December 1998

Contents

Chapter 1

Introduction

1.1 Background

In the first half of this century telephone and telegraph were the main communication medium for the transmission of human voice. The technical progress in communication technologies and microelectronics during the last decades and years led to an increasing significance of radio based mobile communications. Wireless telephones guarantee a high degree of liberty and comfort, a new lifestyle at an achievable price. People can be contacted by telephone wherever they are within the service area of the cellular networks on air. Thus, in 1980 the first standardized mobile terrestrial system called *NMT (Nordic Mobile Telephone)* was introduced [Walke 98]. It was followed in 1985 by *AMPS (Advanced Mobile Phone System)* in USA, *C450 (Cellular)* in Germany and *TACS (Total Access Communication System)* in the UK. Cellular networks can be deployed wherever the installation of cables in the fixed-line telephone network is uneconomic or impossible.

The main limiting factor of analog wireless systems is the available frequency spectrum. Better spectral efficiency can be achieved by digitizing the speech signal and applying modern digital techniques such as modulation, coding and equalization. In this way a bandwidth efficient transmission is possible, better immunity to radio disturbances and less susceptibility to noise than in the case of analog modulated signals. Digitally sampled speech and data signals can be further processed and stored before transmission. This enables the use of different multiplex procedures such as *TDM (Time Division Multiplex)*, *FDM (Frequency Division Multiplex)* or *CDM (Code Division Multiplex)* which guarantee a higher number of subscribers that can be serviced.

Since the introduction of *GSM (Global System for Mobile Communications)* in 1990, an extraordinary increase in digital cellular telephone users has been registered and the trend is still continuing. The worldwide acceptance of the digital technology is reflected in the *Compound Annual Growth Rate (CAGR).*

In the year 2001 (Figure 1.1.a), over 75 % of the market will be using digital technology, dual mode analog/digital units will remain a unique, expensive appendage in the Americas [GSM Data 98]. By 2002 or 2003, analog systems will dissapear as Generation 3 technology (*UMTS*[1], *IMT-2000*[2]) will be rapidly deployed.

Region	CAGR 1997 - 2001
USA	14.97 %
Canada	19.31 %
South America	46.40 %
The Americas	22.20 %
Western Europe	19.19 %
Eastern Europe	41.18 %
Russia	63.46 %
Subtotal Europe	21.54 %
Japan	31.65 %
Rest of the World	55.46 %
Worldwide Total	33.58 %

Table 1.1: Worldwide wireless subscriber growth

Table 1.1 [GSM Data 98] presents the worldwide subscriber forecast by geography as it has been reported by *Cahners In-Stat Group* in January 1998, all 1997 subscriber figures being based on the first half performance in 1997 and then extrapolated for the balance of the year. It can be observed that Eastern Europe, Russia and the rest of the world are expected to exhibit excellent growth rates.

Amongst the mobile communications technologies GSM has proven itself as the most popular [GSM MoU 98a], having more than 100 million customers worldwide. GSM growth had outstripped even the most incautious speculations.

Starting its activity as a purely Pan-European cellular telecommunication standard[3], GSM is now considered to be the de facto global cellular standard. It provides almost complete coverage in Western Europe and growing coverage in Eastern Europe, Asia and the Americas. Roaming is widespread and allows cellular subscribers to use their services in any GSM service area in the world in

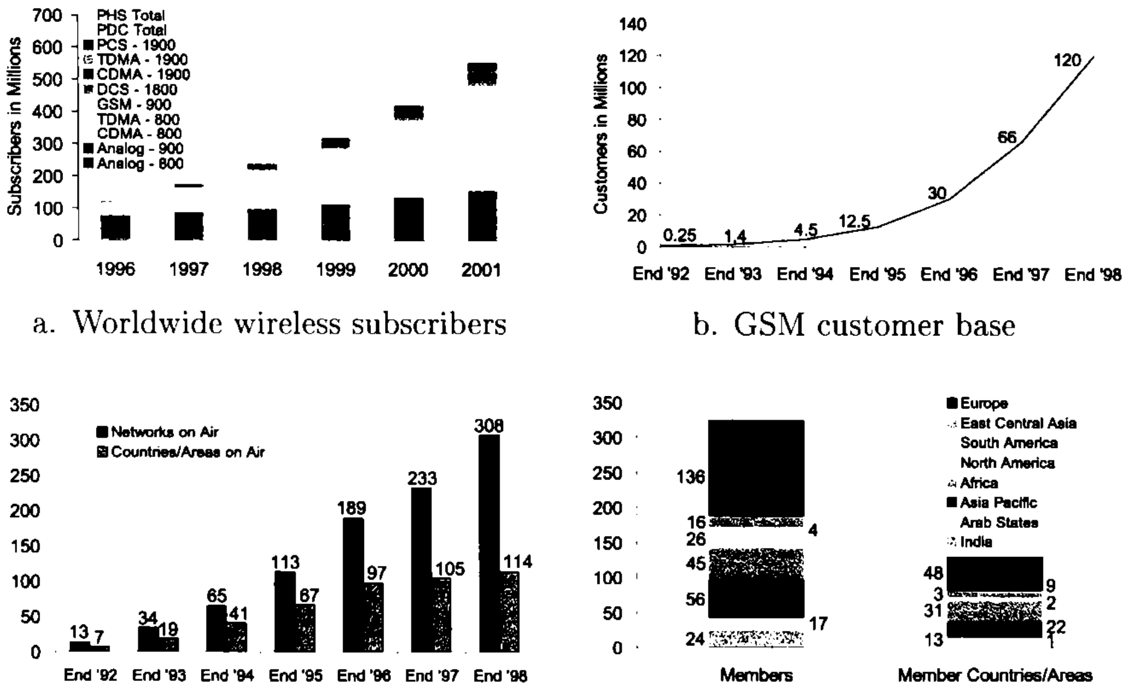
[1] UMTS: Universal Telecommunications System

[2] IMT-2000: International Mobile Telecommunications at 2,000 MHz

[3] According to the *International Standardization Organization (ISO)* a *standard* is a technical specification or other document available to the public, drawn up in cooperation and consensus or general approval of all interests affected by it, based on the consolidated results of science, technology and experience, aimed at the promotion of optimum community benefits and approved by a body recognized on the national, regional or international level [Walke 98].

which their network provider has a roaming agreement. It can be considered that subscribers can virtually go to any country in Europe and some in Asia as well and be assured that their GSM phone will work both for voice and data. For travellers to North America dual-band phones are available which permit the use of both GSM and PCS technology.

Some statistics published by the *GSM MoU Association* are presented in Figure 1.1. The *MoU (Memorandum of Understanding*, Annex B) includes members operating GSM networks at 900 MHz (GSM 900) and at the higher 1,800 MHz (DCS 1800) and 1,900 MHz frequency (PCS).



a. Worldwide wireless subscribers

b. GSM customer base

c. GSM networks and countries/areas on air

d. Distribution of GSM in the world

Figure 1.1: GSM statistics

Figure 1.1.b presents the growth of the GSM customer base since 1992. It can be observed that the number is increasing continuously. The number of GSM networks and countries with GSM on air since 1992 till 1998 is represented in Figure 1.1.c. At the present a mean value of approximately 3 network providers per country is considered.

The worldwide distribution of GSM at the end of 1998 is presented in Figure 1.1.d. GSM is available in 44 % of the world's total cellular market, Germany having the largest number of GSM users. *Mannesmann D2 Privat* is the worldwide biggest GSM network with more than 5 million users. In Germany the number of new subscribers is growing monthly by approximately 500,000 [connect 98]. In October 1998 the number of cellular telephone connections has raised up to approximately 12 millions, the most of them being in the GSM system operating at 900 and 1,800 MHz. The forecast for Germany is a number of 20,000,000 subscribers in the year 2000.

1.2 Scope of This Thesis

Many telephone subscribers use a mobile telephone handset while driving. This jeopardizes driving safety and there are many countries where handsfree operation in car telephones is not only desirable but also mandatory. The goal of a handsfree system is to permit a full-duplex communication with all the comfort specific to conversational speech such as easiness to speak, to listen and to interrupt the remote user, with no restriction on normal gestures and movements. Naturalness and a satisfactory quality of transmitted speech is as important as a minimal, not disturbing acoustic background.

The car handsfree equipment incorporates a microphone and a loudspeaker installed in the car cabin, the mobile terminal being inserted in a cradle. Thus the driver's hands are free to operate the car. The installation of the loudspeaker and microphone inside the vehicle gives arise to acoustic echo problems, because of the reflections from the car interior. The signal picked up by the microphone will therefore consist of an undesired combination of multiple reflections and background noise [Pauler 98]. Two types of impairments will come up when using a car handsfree system, namely the coupling from the loudspeaker to the microphone and the high ambient noise level due to different background sources such as wind, tyre, engine or fan noise. These problems seriously affect the quality of the transmission and degrade speech quality.

Digital communication systems, such as the cellular GSM (Annex B), offer better listening quality than the analog communication networks because of their better immunity to radio disturbances. This better speech quality is achieved at the expense of an increase in transmission times caused by the digital signal processing techniques. The long delays inherent to the GSM system make the acoustic echo problem in handsfree situations much worse than in other lower delay systems. The subscriber on the other end will always hear a disturbing, unacceptable echo, unless echo cancellation is performed in the handsfree equipment.

In handsfree mobile applications, where the acoustic echo is mixed with high background noise, it is recommended to use a combined system of acoustic echo cancellation (AEC) and noise reduction (NR). For a correct elimination of acoustic echo and background noise the combined system must have efficient performances in terms of initial convergence, tracking of the echo path variations, speech enhancement in noisy environment and double-talk situation. The considerable delay introduced by the GSM system will stress any bad or insufficient performance.

The main focus of this thesis lies in the investigation of algorithms used in acoustic echo compensation and noise reduction. These two topics are challenging applications of adaptive filtering in telecommunications. A combined system consisting of an acoustic echo canceller and a noise reduction system suitable for handsfree systems in the GSM network is presented. The proposed system works entirely in the time domain, thus no supplimentary time delay is introduced except the inherent processing time. The correct operation of echo cancellation and noise reduction is enabled by the use of voice activity and double-talk detection algorithms. The algorithms forming the combined system are designed to be implemented in a digital signal processor (DSP).

An introduction to the echo problem, both line and acoustical echo, and the theoretical background of adaptation algorithms are presented in *Chapter 2*. The advantages and disadvantages of these algorithms when used in speech processing are also discussed in this chapter.

An overview of the existing speech enhancement algorithms is presented in *Chapter 3*. The specific noise problem in a vehicle interior is discussed and the goals of speech enhancement in mobile communications are viewed. The presented speech enhancement algorithms consider the "one microphone approach" as well as the adaptive noise cancellation by multiple microphones. Both frequency and time domain algorithms are discussed.

When designing echo cancellation and speech enhancement algorithms the digital model of the vehicle interior must be known. The definition of such a model will be the main task in *Chapter 4* as well as the specification of the GSM requirements for acoustic echo cancellation and noise reduction. In *Chapter 5* the structure of the proposed combined system and the most important objective assessment methods employed during the development phase of speech processing algorithms are presented.

In *Chapter 6* the adaptation algorithm for the combined system is introduced and treated. Different environmental conditions and implementations are considered, the fullband as well as the subband approach. For the fullband performance in noisy near-end environment a new stepsize control algorithm is proposed. As the acoustic echo cancellation and the noise reduction using one microphone cannot be properly executed without any speech detection algorithms, this topic will be accessed in *Chapter 7*. A new voice activity detection algorithm is proposed which will update its adaptive energy threshold depending on its current decision. The combined system will be completed in *Chapter 8* by the presentation of the proposed "one microphone approach" noise reduction system, implemented in the time domain.

A summary of the work will be given in *Chapter 9*.

Chapter 2

Acoustic Echo Cancellation

2.1 Basics of Echo Cancellation

Echoes are defined as being the sound waves reflected from walls, floors or other objects that are arriving short time after the direct sound. If a reflected sound arrives a very short time after the direct sound, it is perceived as reverberation [Sondhi & Kellermann 92]. The desirable amount of this spectral distorsion is application specific: thus, it is the special concern of architects of concert halls, but in the same time it is of no desire in offices. As long as the delay of the reflections does not exceed 40 ms [Taylor 94] the echo is not disturbing.
In telecommunications echoes can be generated electrically and acoustically as well. *Electrical* or *line echoes* are derived from the electrical signal transmitted on the line. *Acoustical echoes* arise in *loudspeaker-room-microphone systems (LRMS)* where the handsfree operation is desirable, e.g. in offices, teleconferencing systems or cars.
The extent of echo annoyance [Goldenberg & Bisson 95] depends on two factors:

- the background noise level, i.e. the circuit noise and/or ambient noise. In a highly noisy environment the echo can merely be perceived, in this case the noise is much more disturbing than the echo.

- the echo delay with respect to the original signal. If the delay is short, it is masked by the original signal, but with increasing delay the subscriber can hear a hollow effect. If the propagation delay reaches several tens of milliseconds, the echo is clearly detached from the original speech and leads to a seriously disturbed conversation.

2.1.1 Line Echoes

The main origin of line echoes is the impedance mismatch at the devices that connect the two wire customer's loop to the four wire circuits. These converters are known as *hybrids* or *hybrid transformers.*

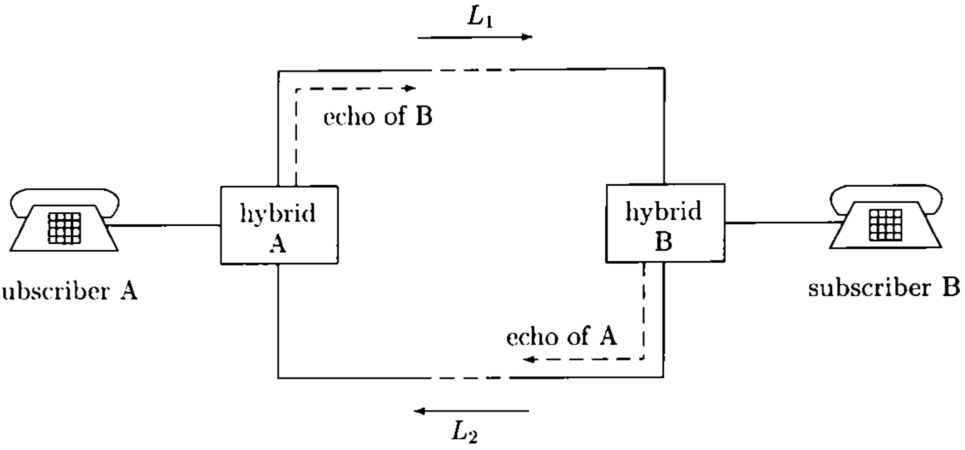


Figure 2.1: Line echo

The hybrid shown in Figure 2.1 is a network bridge, which allows the signals from subscriber A to go along path L_1 to subscriber B and from subscriber B via path L_2 back to A. The signal from A must be prevented from getting back to A along path L_2. The same stands for signals coming from subscriber B, which must be hindered from returning along path L_1. Because there is no possibility of exactly balancing the impedance of the customer's loop, the echo at the hybrid cannot be eliminated completely.
A remedy to this problem are the echo suppressors with voice-operated switches and adaptive line echo cancellers.
The echo suppressor makes its decision based on the level of two speech signals. If the return signal is low in level, it is considered to be only echo and the return path is opened. Alternatively, when the level in the return path is high, this signal is considered to be an interruption, the other subscriber trying to break into the conversation. These suppressors work well on circuits with a roundtrip-delay of up to 100 ms.
With long roundtrip-delays, as in telephone communications via satellite, echo suppressors fail to work correctly. In this case, adaptive echo cancellers are used, which subtract a synthetically generated echo from the return signal. This synthetic signal is generated by passing the speech signal through a filter whose impulse response matches the echo path (Figure 2.2).

The adaptive echo cancellation can be considered as being a system identification problem [Stearns & David 93]: the unknown echo path has to be modelled by a filter. Because of the changing echo path this filter has to be made adaptive, thus being able to track slow variations of the characteristics of the subscriber's loop. After a fast initial convergence, the tracking mode can be slowed

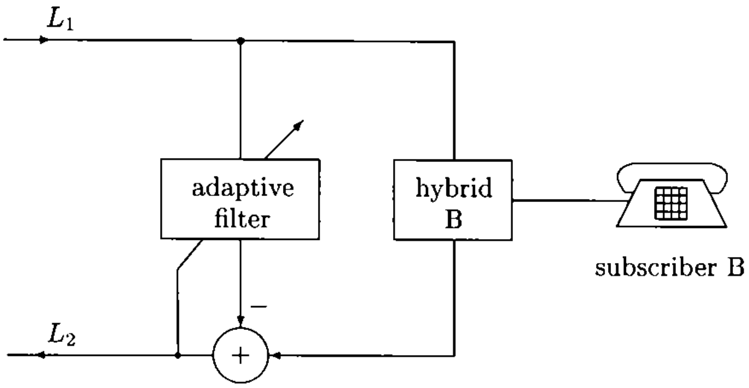


Figure 2.2: Adaptive line echo canceller

down, as the line echo path is assumed to change very slowly.
Concerning the circuit and background noise, the levels are generally low and do not cause any problems in the line echo cancellation system.

2.1.2 Acoustic Echoes

A fundamental problem of conventional handsfree equipment is the coupling from the louspeaker to the microphone. The acoustic feedback exists due to the direct propagation from the loudspeaker to the microphone and the indirect sound spreading of the reflections from surrounding objects in the room. The microphone is located some significant distance from the speaker's mouth and in the sound field of the loudspeaker of the handsfree system [Naylor et al. 94], [Taylor 94]. These coupling effects feed back the far-end speech into the microphone, so that the far-end subscriber will hear himself after a network-dependent delay time. This very annoying effect is called acoustic echo. In many cases this feedback can also lead to instabilities which result in howling.
Solving properly the acoustic echo control problem is crucial for the design of audio terminals for telecommunications operating in handsfree mode, such as handsfree telephone sets, teleconference systems and videoconferencing devices. Adaptive echo cancellers, which construct a replica of the actual acoustic echo path by means of an adaptive filtering method, are recognized as the best solution to the acoustic echo problem [Gilloire 95].

Acoustic echoes are usually much more delayed than line echoes. Therefore the computational and memory requirements are much higher for acoustic echo cancellers than for line echo cancellers [Weiss et al. 91]. The acoustic echo path is usually several times longer and it may change rapidly, at any time, due to moving persons in the room, opening windows or doors, temperature changes, etc. Thus, the acoustic echo canceller has to compensate longer impulse responses and it also has to converge faster than the line echo canceller [Sondhi & Kellermann 92].

Acoustic echo cancellation is a system identification problem dealing with long, nonstationary impulse responses and high nonstationary noise at the observable output.

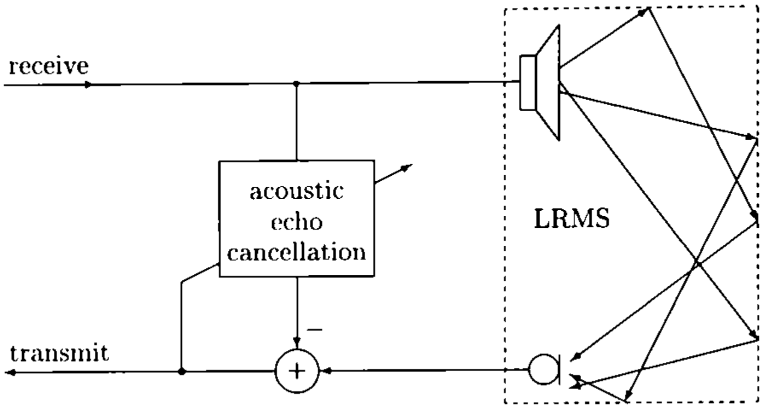


Figure 2.3: Adaptive acoustic echo canceller

The acoustic echo canceller reproduces the acoustic room impulse response, i.e. the transmission characteristics between the loudspeaker and the microphone. It creates a synthetic echo signal which eliminates the room echo by subtraction (Figure 2.3). The performance of the echo canceller depends on the adaptation algorithm and the length of the adaptive filter. The filter coefficients must constantly be updated since the echo path is changing even by the smallest movements of people in the room. For a transversal finite impulse response (FIR) filter[1] the number of coefficients L is dependent on the reverberation time t_N of the room and the sampling frequency f_S of the digital system. The more coefficients the filter has, the longer are the echoes that can be estimated and compensated:

$$L = t_N \cdot f_S \tag{2.1}$$

If the FIR-filter has to achieve an attenuation of 60 dB it needs to be as long as the reverberation time[2].

The reverberation time of a room is given [Frenzel 92] by the following formula:

$$t_N = \frac{24\,V \ln 10}{c\,S\,\bar{\alpha}} \quad [\mathrm{s}] \tag{2.2}$$

where V represents the volume of the considered room in m^3, c is the sound speed in m/s and $\bar{\alpha}$ is called the absorption coefficient

$$\bar{\alpha} = \frac{1}{S}\sum_i S_i\,\alpha_i \tag{2.3}$$

[1] See section 4.1.

[2] Every room can be characterized by its reverberation time, defined as the time the sound energy needs to decrease by 60 dB after switching off the sound event [Armbruster et al. 91].

S_i are the surface areas in m^2 with different absorption. Table 2.1 includes some absorption coefficients at a frequency of 1 kHz [Ansahl 98], [Kuttruff 91], [Schneider 94], [Siemens AG 97].

Material	α
Glass	0,02
PVC, Rubber on floor	0,02
Steel	0,03
Plywood, 3 mm thickness	0,09
Suede carpet	0,20
Plastic (Polyurethan)	0,70
Person on upholstered chair	0,87

Table 2.1: Acoustic absorption coefficients α of some materials

The reverberation time also depends on temperature changes in the room. The sound speed in equation (2.2) is temperature-dependent [Kuttruff 91]:

$$c = (331.4 + 0.6\ \vartheta) \quad \left[\frac{\mathrm{m}}{\mathrm{s}}\right] \tag{2.4}$$

where ϑ represents the room temperature in °C.

The degree of absorption also depends on the surface structure of the absorbent material. Thus, porous materials usually absorb high frequencies much better than low frequencies. Likewise, high frequencies are less absorbed by oscillating, vibrating surfaces [Siemens AG 97].

Taking into account the above statements, one can differenciate between acoustic echo cancellers for two different application situations: an office environment when performing an audioconference or a vehicle interior when using the handsfree function of a mobile phone. Offices are much larger than cars, therefore the length of the impulse response of an office is often hundreds of milliseconds, while in a vehicle the environment is almost ideal with respect to reverberation time. The small internal volume and the upholstery of the interior lead to reverberation times of less than 60 ms. This corresponds to an FIR-filter length of 480 coefficients, when sampling with 8,000 Hz.
In a car environment a second impairment exists, namely the omnipresent background noise. The perceived effect of this additive noise is listening fatigue and a reduction in speech intelligibility. Acoustic echo cancellers usually have difficulty in processing the echo in high background noise.

2.2 Adaptation Algorithms

An adaptive algorithm is the equation or the set of equations used to adjust the transfer function of the adaptive system. The elements of a basic adaptive system are shown in Figure 2.4.

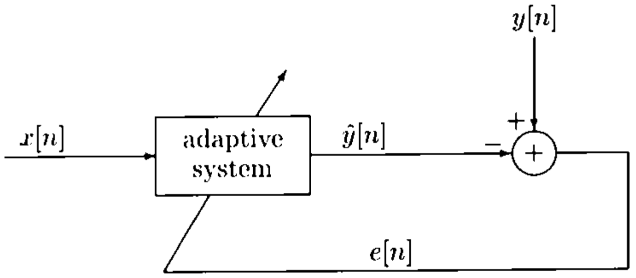


Figure 2.4: Elements of the basic adaptive system

$x[n]$ and $\hat{y}[n]$ are the input and output signals of the adaptive system and $y[n]$ is the desired response. The characteristics of the adaptive system change, or *adapt*, according to signal conditions, so that the error signal $e[n]$ should be minimal. Adaptive signal processing is dealing with time-varying digital systems, n denotes a certain time instant.

An adaptive filter is a selfdesigning device [Haykin 96], in that it relies on a recursive algorithm, which makes it possible for the filter to perform satisfactorily in an environment where complete knowledge of the relevant signal characteristics is not available. The algorithm starts from some predetermined set of initial conditions and, in a stationary environment, it converges to the desired signal. In a nonstationary environment, the algorithm offers a tracking capability, which means that it can track time variations in the statistics of the input data, provided that the variations are sufficiently slow.

If the adaptive filter is to be an accurate model of the unknown system, two conditions must be satisfied [Stearns & David 93]:

- the input signal $x[n]$ must contain enough information to excite all modes of the system. For off-line system identification white noise is a viable solution, but for real-time applications, the selection of the input signal is important.

- the adaptive system must be of sufficient complexity, i.e. high enough in order, to match the degrees of freedom of the unknown system. If an FIR model is used to identify an IIR system, typically a very high order system is required. If the model order is overspecified, the adaptive process may converge more slowly.

The operation of a linear adaptive filtering algorithm involves two basic processes:

- a filtering process designed to produce an output in response to a sequence of input data
- an adaptation process, the purpose of which is to provide a mechanism for the adaptive control of a set of parameters used in the filtering process

These two processes work interactively with each other.

2.2.1 Principle of Orthogonality

The mathematical solution to the problem of minimizing the estimation error $e[n]$ leads to the derivation of a very important theorem, namely the the principle of orthogonality. The following filtering problem will be considered: the filter input is denoted by the time series $x[0], x[1], x[2], \ldots$ and the impulse response of the filter is denoted by $h_0, h_1, h_2, \ldots$. Both time series are assumed to have infinite duration. The output of the device at discrete time n is defined by the linear convolution sum:

$$\hat{y}[n] = \sum_{k=0}^{\infty} h_k x[n-k] \qquad n = 0, 1, 2, \ldots \tag{2.5}$$

The purpose of the filter is to produce an estimate of the desired response denoted by $y[n]$. The estimation of $y[n]$ is accompanied by an error $e[n]$, defined by the difference

$$e[n] = y[n] - \hat{y}[n]. \tag{2.6}$$

In the Ensemble Sense

Assuming that the filter input and the desired output are wide-sense stationary stochastic processes, the *mean-square value of the estimation error e[n]* will be chosen for the optimization of the filter design. The cost function[3] will thus be

$$J = E[|e[n]|^2] \tag{2.7}$$

where E denotes the statistical expectation operator.
Applying the gradient operator ∇ to the cost function J, with respect to the k-th filter coefficient, and setting the gradient vector $\nabla_k J$ to zero will lead to the specification of the operating conditions for which J attains a minimum value.

[3] A cost function provides a quantitative measure for assessing the quality of performance. The cost function is a scalar.

Denoting by $e_o[n]$ the special value of the estimation error that results when the filter operates at its optimum condition, it can be written

$$E[x[n-k]e_o[n]] = 0 \qquad k = 0, 1, 2, \ldots \tag{2.8}$$

Equation (2.8) states the *principle of orthogonality:*
The necessary and sufficient condition for the cost function J to attain its minimum value is that the corresponding value of the estimation error $e_o[n]$ is orthogonal to each input sample that enters into the estimation of the desired response at time n [Haykin 96].

Substituting (2.5) and (2.6) in (2.8), expanding and rearranging the terms will lead to

$$\sum_{i=0}^{\infty} h_{oi} E[x[n-k]x[n-i]] = E[x[n-k]y[n]] \qquad k = 0, 1, 2, \ldots \tag{2.9}$$

where the two expectations represent the following:

- $E[x[n-k]x[n-i]]$ is the *autocorrelation function* of the filter input for a lag of $i-k$ defined as

$$r[i-k] = E[x[n-k]x[n-i]] \tag{2.10}$$

- $E[x[n-k]y[n]]$ is the *crosscorrelation function* between the filter input and the desired response for a lag of $-k$ defined as

$$p[-k] = E[x[n-k]y[n]] \tag{2.11}$$

Using (2.10) and (2.11) in (2.9) an infinitely large system of equations will be found, the so-called the *Wiener-Hopf equations*

$$\sum_{i=0}^{\infty} h_{oi} r[i-k] = p[-k] \qquad k = 0, 1, 2, \ldots \tag{2.12}$$

The system (2.12) defines the optimum filter coefficients in terms of two correlation functions: the autocorrelation function of the filter input and the crosscorrelation function between the filter input and the desired response.

Considering the special case of the linear transversal filter of length L in Figure 2.5 the Wiener-Hopf equations (2.12) reduce to a system of L simultaneous equations:

$$\sum_{i=0}^{L-1} h_{oi} r[i-k] = p[-k] \qquad k = 0, 1, 2, \ldots, L-1 \tag{2.13}$$

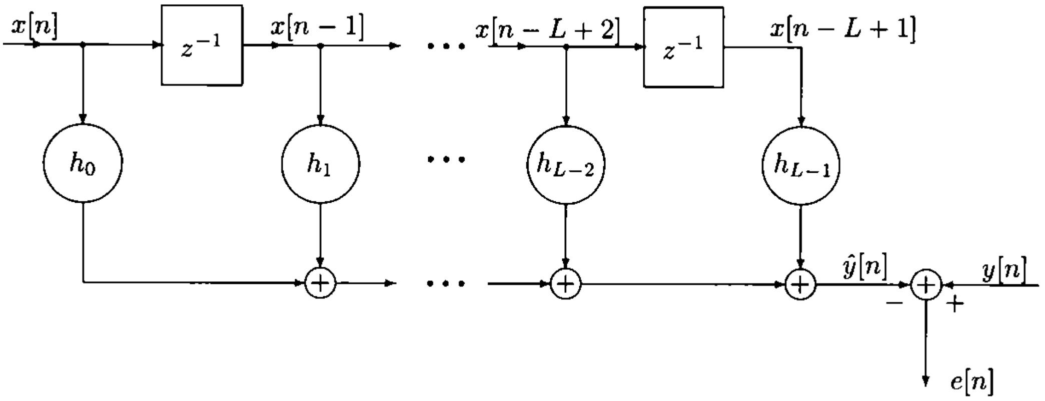


Figure 2.5: Transversal filter

where $h_{o0}, h_{o1}, h_{o2}, \ldots, h_{oL}$ are the optimum values of the tap weights of the filter. The Wiener-Hopf equations provide the mathematical basis of the class of *linear optimum discrete-time filters* also known as *Wiener filters.*

The Wiener-Hopf equations can also be written in compact matrix form:

$$\mathbf{R}\mathbf{h}_o = \mathbf{p} \tag{2.14}$$

where $\mathbf{R}$ represents the $(L \times L)$ *correlation matrix* of the tap inputs $x[n], x[n-1], \ldots, x[n-L+1]$:

$$\mathbf{R} = E[\mathbf{x}[n]\mathbf{x}^T[n]] \tag{2.15}$$

with

$$\mathbf{x}[n] = [x[n], x[n-1], \ldots, x[n-L+1]]^T \tag{2.16}$$

or in expanded form

$$\mathbf{R} = \begin{bmatrix} r[0] & r[1] & \cdots & r[L-1] \\ r[1] & r[0] & \cdots & r[L-2] \\ \vdots & \vdots & & \vdots \\ r[L-1] & r[L-2] & & r[0] \end{bmatrix} \tag{2.17}$$

Correspondingly, $\mathbf{p}$ represents the *$(L\times 1)$ crosscorrelation vector* between the tap inputs of the filter and the desired response *$y[n]$*:

$$\mathbf{p} = E[\mathbf{x}[n]y[n]] \tag{2.18}$$

or in expanded form

$$\mathbf{p} = [p[0], p[-1], \ldots, p[1-L]]^T \tag{2.19}$$

The *$(L\times 1)$ optimum tap-weight vector* $\mathbf{h}_o$ of the transversal filter is

$$\mathbf{h}_o = [h_{o0}, h_{o1}, \ldots, h_{oL-1}]^T \tag{2.20}$$

Assuming $\mathbf{R}^{-1}$ exists, i.e. $\mathbf{R}$ is nonsingular, Eq. (2.14) may be solved for the optimum tap-weight vector:

$$\mathbf{h_o} = \mathbf{R}^{-1}\mathbf{p} \tag{2.21}$$

Equation (2.21) states that the optimum tap-weight vector $\mathbf{h_o}$ is uniquely defined by the product of the inverse of the correlation matrix $\mathbf{R}$ and the cross-correlation $\mathbf{p}$ between the tap input vector $\mathbf{x}[n]$ and the desired response $y[n]$.

Based on Time Average

There are no assumptions made on the statistics of the input of the transversal filter. To optimize the filter design, the *sum of error squares* will be chosen. The cost function will thus be

$$\mathcal{E}(h_0, h_1, \ldots, h_{L-1}) = \sum_{n=n_1}^{n_2} |e[n]|^2 \tag{2.22}$$

where n_1 and n_2 define the index limits at which the error minimization occurs. This sum may also be viewed as an *error energy* [Haykin 96]. For the minimization, the tap weights of the transversal filter $h_0, h_1, \ldots, h_{L-1}$ are held constant during the interval $n_1 \leq n \leq n_2$. To make sure that all the L tap inputs of the transversal filter have nonzero values the limits $n_1 = L$ and $n_2 = N$ will be chosen.

Applying the *gradient operator* ∇ to the cost function $\mathcal{E}$, with respect to the k-th filter coefficient, and setting the gradient vector $\nabla_k\mathcal{E}$ to zero will lead to the specification of the operating conditions for which $\mathcal{E}$ attains a minimum value. Denoting by $e_{min}[n]$ the special value of the estimation error that results when the transversal filter is optimized, it can be written as

$$\sum_{n=L}^{N} x[n-k]e_{min}[n] = 0 \qquad k = 0, 1, \ldots, L-1 \tag{2.23}$$

Equation (2.23) is the mathematical description of the temporal version of the *principle of orthogonality:*

The minimum error time series $e_{min}[n]$ is orthogonal to the time series $x[n-k]$ applied to tap k of a transversal filter of length L for $k = 0, 1, \ldots, L-1$ when the filter is operating in its least-square condition [Haykin 96].

The filter resulting from the minimization is called a *linear least-squares filter*. Proceeding in a similar way to the derivation of the Wiener-Hopf equations, the *system of the normal equations* of a linear least-squares filter will be

$$\sum_{t=0}^{L-1} \hat{h}_t \, \phi[t,k] = z[-k] \qquad k = 0, 1, 2, \ldots, L-1 \tag{2.24}$$

where $\hat{h}_t$ are the special values of the tap-weights for an optimized transversal filter. $\phi[t,k]$ and $z[-k]$ have the following meanings:

- $\phi[t,k]$ represents the *time averaged autocorrelation function (over i)* of the tap inputs of the transversal filter in Figure 2.5

$$\phi[t,k] = \sum_{i=L}^{N} x[i-k]x[i-t] \qquad 0 \leq [t,k] \leq L-1 \tag{2.25}$$

- $z[-k]$ represents the *time averaged crosscorrelation (also over i)* between the tap inputs and the desired response

$$z[-k] = \sum_{i=L}^{N} x[i-k]y[i] \qquad 0 \leq k \leq L-1 \tag{2.26}$$

The normal equations for linear least squares filters can also be written in compact *matrix form:*

$$\mathbf{\Phi}\hat{\mathbf{h}} = \mathbf{z} \tag{2.27}$$

where $\mathbf{\Phi}$ represents the $(L \times L)$ *time averaged correlation matrix* of the tap inputs $x[n], x[n-1], \ldots, x[n-L+1]$

$$\mathbf{\Phi} = \begin{bmatrix} \phi[0,0] & \phi[1,0] & & \phi[L-1,0] \\ \phi[0,1] & \phi[1,1] & & \phi[L-1,1] \\ \vdots & \vdots & & \vdots \\ \phi[0,L-1] & \phi[1,L-1] & \cdots & \phi[L-1,L-1] \end{bmatrix} \tag{2.28}$$

Correspondingly, $\mathbf{z}$ represents the *(L×1) time averaged crosscorrelation vector* between the tap inputs of the filter and the desired response *y[n]*:

$$\mathbf{z} = [z[0], z[-1], \ldots, z[1-L]]^T \tag{2.29}$$

The *(L×1) tap-weight vector of least squares filter* $\hat{\mathbf{h}}$ of the transversal filter is

$$\hat{\mathbf{h}} = [\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_{L-1}]^T \tag{2.30}$$

Assuming $\mathbf{\Phi}^{-1}$ exists, i.e. $\mathbf{\Phi}$ is nonsingular, Eq. (2.27) may be solved for the tap-weight vector of the least-squares filter:

$$\hat{\mathbf{h}} = \mathbf{\Phi}^{-1}\mathbf{z} \tag{2.31}$$

This last equation is fundamental to the recursive formulations of the linear least-squares filter. Equation (2.31) states that the tap-weight vector of the least-squares filter $\hat{\mathbf{h}}$ is uniquely defined by the product of the inverse of the time averaged correlation matrix $\mathbf{\Phi}$ and the time averaged crosscorrelation $\mathbf{z}$ between the tap input vector $\mathbf{x}[n]$ and the desired response $y[n]$.

The time averaged correlation matrix $\mathbf{\Phi}$ can be approximated [Haykin 96] as

$$\mathbf{\Phi}[n] \simeq \frac{\mathbf{R}}{1 - \lambda} \qquad n \text{ large} \tag{2.32}$$

and in a corresponding way, the inverse matrix $\mathbf{\Phi}^{-1}$ may be expressed as

$$\mathbf{\Phi}^{-1}[n] \simeq (1 - \lambda)\,\mathbf{R}^{-1} \qquad n \text{ large} \tag{2.33}$$

where $\mathbf{R}^{-1}$ is the inverse of the ensemble-averaged correlation matrix $\mathbf{R}$ and λ is a positive constant close to, but less than, 1.

Equation (2.31) is the least-squares counterpart to the solution of the matrix form of the Wiener-Hopf equations (2.21).

2.2.2 Least Mean Squares (LMS)

The requirement of an adaptive transversal filter is to modify the tap-weight vector $\mathbf{h}$ in the direction of the optimum tap-weight vector $\mathbf{h_o}$, i.e. to satisfy, after adaption, the Wiener-Hopf equations (2.14).

The LMS adaptation method is defined [DeGroat et al. 97] by the following recursiv relation:

$$\mathbf{h}[n+1] = \mathbf{h}[n] + \frac{1}{2}\,\mu\,(\Delta\mathbf{h}[n]) \tag{2.34}$$

where $\mathbf{h}[n+1]$ represents the updated tap-weight vector at time instant $n+1$, μ is a positive real-valued constant and $\Delta\mathbf{h}$, the *adjustment vector* is given by

$$\Delta\mathbf{h}[n] = -\nabla_h J[n]. \tag{2.35}$$

$\nabla_h J$ represents the gradient vector with respect to $\mathbf{h}[n]$ at time instant n, the factor $\frac{1}{2}$ is used merely for the purpose of cancelling a factor 2 from the gradient vector formula.

As exact measurements of the gradient vector are not possible, the gradient vector must be estimated from the available data. The stochastic gradient algorithm replaces the *expected value* of the mean squared error by its *instantaneous value* [Sondhi & Kellermann 92]. Thus, the stochastic gradient version of Eq. (2.7) will be

$$J[n] = e[n]^2 \tag{2.36}$$

For the LMS algorithm being a representative of the stochastic gradient algorithms, the estimated adjustment vector $\Delta\hat{\mathbf{h}}[n]$ can be derived [Vaseghi 96] as follows:

$$\begin{aligned} -\hat{\nabla}_h J[n] &= -\nabla_{\hat{h}}\left[e^2[n]\right] \\ &= -\nabla_{\hat{h}}(y[n] - \hat{\mathbf{h}}[n]^T\mathbf{x}[n])^2 \\ &= 2\mathbf{x}[n](y[n] - \hat{\mathbf{h}}[n]^T\mathbf{x}[n]) \\ &= 2\mathbf{x}[n]e[n] \end{aligned} \tag{2.37}$$

where $\hat{\nabla}_h J$ represents the instantaneous estimate of the gradient vector. After substituting (2.37) in the recursive relation (2.34) the LMS adaptation equation is found:

$$\hat{\mathbf{h}}[n+1] = \hat{\mathbf{h}}[n] + \mu\mathbf{x}[n]e[n]. \tag{2.38}$$

This result may be written in the form of three basic relations [Haykin 96]:

1. *filter output:*
$$\hat{y}[n] = \hat{\mathbf{h}}[n]^T\mathbf{x}[n] \tag{2.39}$$

2. *estimation error:*
$$e[n] = y[n] - \hat{y}[n] \tag{2.40}$$

3. *tap-weight adaptation:*
$$\hat{\mathbf{h}}[n+1] = \hat{\mathbf{h}}[n] + \mu\,\mathbf{x}[n]e[n] \tag{2.41}$$

Convergence Rate

The convergence behaviour of the LMS algorithm depends on two major factors:

- the stepsize parameter μ
- the eigenvalues of the correlation matrix $\mathbf{R}$ of the tap-input vector $\mathbf{x}[n]$.

For the LMS algorithm to be convergent in the mean square, it is neccesarry that the stepsize parameter μ satisfies the condition [Widrow et al. 75]

$$0 < \mu < \frac{2}{\lambda_{max}} \tag{2.42}$$

where λ_{max} is the largest eigenvalue of the correlation matrix $\mathbf{R}$. As λ_{max} is generally not available, but knowing that

$$\lambda_{max} \leq \ tr\,[\mathbf{R}] \tag{2.43}$$

where $tr[\mathbf{R}]$ denotes the trace of the matrix $\mathbf{R}$, the condition (2.42) may be reformulated as

$$0 < \mu < \frac{2}{tr\,[\mathbf{R}]} \tag{2.44}$$

Because the correlation matrix $\mathbf{R}$ is Toeplitz, with all of the elements on its main diagonal equal to $r[0]$ it can be written

$$\begin{aligned} tr[\mathbf{R}] &= L \cdot r[0] \\ &= \sum_{k=0}^{L-1} E[|\,x[n-k]\,|^2] \\ &= \text{tap-input power} \end{aligned} \tag{2.45}$$

This means that the condition of Eq. (2.42) can be expressed as

$$0 < \mu < \frac{2}{\text{tap-input power}} \tag{2.46}$$

The *misadjustment* $\mathcal{M}$, a dimensionless quantity providing a measure of how close the LMS algorithm (after completed adaptation) is to the optimum in the mean-squared error sense, can be aproximated [Haykin 96] as follows:

$$\begin{aligned} \mathcal{M} &= \frac{\mu}{2} \sum_{i=1}^{L} \lambda_i \\ &= \frac{\mu}{2} \cdot (\text{tap-input power}) \end{aligned} \tag{2.47}$$

The smaller $\mathcal{M}$, the closer the adaptive filtering operation of the LMS algorithm is to optimality.

The *mean convergence rate*, or *exponential convergence time constant* τ_k along the kth eigenvector λ_k of the correlation matrix $\mathbf{R}$, can be approximated as [Widrow & Stearns 85]

$$\tau_k \approx \frac{1}{2\,\mu\,\lambda_k} \tag{2.48}$$

It can be seen that the convergence rate is a function of both the stepsize μ and the spread of eigenvalues $\lambda_{max}/\lambda_{min}$ of the correlation matrix $\mathbf{R}$. The smaller μ, the slower the adaptation. For a large condition number[4] the convergence time may be quite slow.

Defining an *average eigenvalue* for $\mathbf{R}$

$$\lambda_{av} = \frac{1}{L} \sum_{i=1}^{L} \lambda_i \tag{2.49}$$

and an *average time constant* τ_{av}

$$\tau_{av} \approx \frac{1}{2\,\mu\,\lambda_{av}} \tag{2.50}$$

the relation that exists between the misadjustment $\mathcal{M}$ and the average time constant τ_{av} can be written as

$$\begin{aligned} \mathcal{M} &\approx \frac{\mu\,L\,\lambda_{av}}{2} \\ &\approx \frac{L}{4\,\tau_{av}} \end{aligned} \tag{2.51}$$

From this formula it can be observed that

[4] The condition number of a Hermitian matrix is defined as the ratio of its largest eigenvalue to its smallest eigenvalue

- the misadjustment $\mathcal{M}$ increases linearly with the number of taps L for a fixed τ_{av}
- the misadjustment $\mathcal{M}$ is directly proportional to the parameter μ, whereas the average time constant τ_{av} is inversely proportional to μ.

Therefore, selecting the value of the stepsize parameter μ involves a trade-off between the rate of convergence and the accuracy of the adaptive filtering. When μ is assigned a small value, the misadjustment after adaptation is small but the adaptation is slow. On the other hand, choosing a large μ, the adaptation is relatively fast, but at the expense of an increase of $\mathcal{M}$.

Tracking Capability

The LMS algorithm is *model-independent* and therefore has good tracking performance in a nonstationary environment, provided the statistical variations of the input data are sufficiently slowly varying with respect to the convergence rate of the algorithm.
The LMS algorithm continuously tracks the minimum point of the error performance surface[5].

Computational Load

Due to its simplicity, the LMS algorithm requires a number of computations (multiplies/adds) per time update proportional to the number of adjustable weights L in the algorithm. The LMS algorithm needs only $2L$ operations per time update [Marple 87].

Normalized Least Mean Square (NLMS)

The adjustment vector $\Delta\hat{\mathbf{h}}[n]$ applied to the tap-weight vector $\hat{\mathbf{h}}[n]$ at iteration $i+1$ is directly proportional to the input vector $\mathbf{x}[n]$, which means that for large input values the error signal will have great influence on the filter coefficient update. This effect is known as *gradient noise amplification.* To overcome this difficulty in the adaptation process, the correction applied to the tap-weight vector $\hat{\mathbf{h}}[n]$ at iteration $i+1$ will be normalized with respect to the squared Euclidean norm of the tap-input vector

$$\Delta\hat{\mathbf{h}}[n] = \frac{\tilde{\mu}}{\|\mathbf{x}[n]\|^2}\mathbf{x}[n]\,e[n] \tag{2.52}$$

[5] The dependence of the mean-squared error on the unknown tap weights is referred to as the *error-performance surface.* The tap weights corresponding to the minimum point of this surface define the optimum Wiener solution [Haykin 96].

The adaptation constant $\tilde{\mu}$ for the NLMS is thus dimensionless, whereas the adaptation constant μ has the dimensions of inverse power

$$\mu[n] = \frac{\tilde{\mu}}{\| \mathbf{x}[n] \|^2} \tag{2.53}$$

Considering this statement, the NLMS algorithm can be viewed as an LMS algorithm with a time-varying stepsize parameter.
For convergence in the mean square, the NLMS adaptation constant $\tilde{\mu}$ must satisfy the condition

$$0 < \tilde{\mu} < 2 \tag{2.54}$$

The NLMS algorithm exhibits a faster convergence rate than that of the standard LMS algorithm for both correlated and uncorrelated input data [Haykin 96]. By overcoming the difficulty of gradient noise amplification, the NLMS algorithm introduces a problem of its own, namely numerical difficulties for very small input signal because of the division by a very small value for the squared norm $\| \mathbf{x}[n] \|^2$.

2.2.3 Recursive Least Squares (RLS)

Based on the method of least squares, the *least-squares (LS) algorithm* and its recursive version and the sample-by-sample *recursive least-squares (RLS) algorithm* can be determined.

The LS algorithm is a block processing algorithm. An optimum estimate of the tap-weight vector $\hat{\mathbf{h}}[n]$ is derived from a block of data. This estimate is assumed to be valid until the next block of data is processed to give a new estimate of $\hat{\mathbf{h}}[n]$ [Sondhi & Kellermann 92]. The mean drawbacks of block processing algorithms are the delay they introduce and their lower tracking capability in nonstationary environments [Vaseghi 96].

An alternative to the LS algorithm is the RLS algorithm, in which an optimal estimate of $\hat{\mathbf{h}}[n]$ is obtained recursively at every time instant. In the recursive implementation of the least squares method, the adaptation starts with known initial conditions and the information contained in the new data samples is used to update the old estimate [Haykin 96]. At every time instant, the estimated tap-weight vector minimizes the following cost function

$$\mathcal{E}[n] = \sum_{i=1}^{n} \lambda^{n-i} \, | e[i] |^2 \tag{2.55}$$

with λ^{n-i} being the *exponential weighting factor* or *forgetting factor*. λ is a positive constant in the range $0 < \lambda < 1$. By the use of the forgetting it can be ensured, that data in the distant past do not affect the current estimate. The

"forgetting" of remote data is important when the filter operates in nonstationary environment. The inverse of $1 - \lambda$ can be considered as the *memory* of the algorithm. $\mathcal{E}[n]$ is depending on the variable length n of data.
Based on the *normal equations*, the optimum value of the tap-weight vector $\hat{\mathbf{h}}[n]$, for which the cost function of Eq. (2.55) attains its minimum is

$$\hat{\mathbf{h}}[n] = \mathbf{\Phi}[n]^{-1}\mathbf{z}[n] \tag{2.56}$$

where [Haykin 96]

$$\mathbf{\Phi}[n] = \sum_{i=1}^{n} \lambda^{n-i}\,\mathbf{x}[i]\,\mathbf{x}^T[i] \tag{2.57}$$

and

$$\mathbf{z}[n] = \sum_{i=1}^{n} \lambda^{n-i}\,\mathbf{x}[i]y[i] \tag{2.58}$$

Isolating the term corresponding to $i = n$ from the summation in Eq. (2.57), it may be written

$$\mathbf{\Phi}[n] = \lambda\,\mathbf{\Phi}[n-1] + \mathbf{x}[n]\,\mathbf{x}^T[n] \tag{2.59}$$

where $\mathbf{\Phi}[n-1]$ is the value of the correlation matrix at time $n-1$ and the matrix product $\mathbf{x}[n]\,\mathbf{x}^T[n]$ is the update term in the recursive operation.
Similarly, the crosscorrelation vector $\mathbf{z}[n]$ between the tap inputs and the desired response is

$$\mathbf{z}[n] = \lambda\,\mathbf{z}[n-1] + \mathbf{x}[n]y[n]. \tag{2.60}$$

For computing the least-square estimate $\hat{\mathbf{h}}$ the inverse of the correlation matrix $\mathbf{\Phi}[n]$ is needed. Using the *matrix inversion lemma* or *Woodbury's identity* known from matrix algebra, a recursive implementation for the inverse of the correlation matrix $\mathbf{\Phi}[n]$ can be obtained.
The matrix inversion lemma states that, if there is given a matrix $\mathbf{A}$ defined as

$$\mathbf{A} = \mathbf{B}^{-1} + \mathbf{C}\mathbf{D}^{-1}\mathbf{C}^T \tag{2.61}$$

then its inverse $\mathbf{A}^{-1}$ can be determined by using the following equation:

$$\mathbf{A}^{-1} = \mathbf{B} - \mathbf{B}\mathbf{C}(\mathbf{D} + \mathbf{C}^T\mathbf{B}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{B}. \tag{2.62}$$

where $\mathbf{A}$ and $\mathbf{B}$ are two positive-definite $(L \times L)$ matrices, $\mathbf{D}$ is also a positive-definite matrix of dimension $(N \times N)$ and $\mathbf{C}$ is a $(L \times N)$ matrix [Vaseghi 96].

Assuming that $\mathbf{\Phi}[n]$ is positive-definite and therefore nonsingular, the following identifications can be considered:

$$\mathbf{A} = \mathbf{\Phi}[n] \tag{2.63}$$

$$\mathbf{B} = \lambda\mathbf{\Phi}[n-1] \tag{2.64}$$

$$\mathbf{C} = \mathbf{x}[n] \tag{2.65}$$

$$\mathbf{D} = 1 \tag{2.66}$$

Substituting Eqs. (2.63 - 2.66) in Eq. (2.62) leads to the following recursive equation for the inverse of the correlation matrix:

$$\mathbf{\Phi}^{-1}[n] = \lambda^{-1}\,\mathbf{\Phi}^{-1}[n-1] - \frac{\lambda^{-2}\,\mathbf{\Phi}^{-1}[n-1]\,\mathbf{x}[n]\,\mathbf{x}^{T}[n]\,\mathbf{\Phi}^{-1}[n-1]}{1+\lambda^{-1}\mathbf{x}^{T}[n]\,\mathbf{\Phi}^{-1}[n-1]\,\mathbf{x}[n]} \tag{2.67}$$

Defining the variables $\mathbf{P}[n]$ and $\mathbf{k}[n]$ as

$$\mathbf{P}[n] = \mathbf{\Phi}^{-1}[n] \tag{2.68}$$

and

$$\mathbf{k}[n] = \frac{\lambda^{-1}\,\mathbf{P}[n-1]\,\mathbf{x}[n]}{1+\lambda^{-1}\mathbf{x}^{T}[n]\,\mathbf{P}[n-1]\,\mathbf{x}[n]}, \tag{2.69}$$

equation (2.67) may be rewritten

$$\mathbf{P}[n] = \lambda^{-1}\,\mathbf{P}[n-1] - \lambda^{-1}\,\mathbf{k}[n]\,\mathbf{x}^{T}[n]\,\mathbf{P}[n-1] \tag{2.70}$$

Equation (2.70) is known as the *Riccati equation* for the RLS algorithm. $\mathbf{P}[n]$ is referred to as the *inverse correlation matrix* and $\mathbf{k}[n]$ as the *gain vector*. The time update for the tap-weight vector $\hat{\mathbf{h}}$ is derived from the normal equation (2.27) using the definitions for $\mathbf{k}[n]$ and $\mathbf{P}[n]$, the recursive update operation for $\mathbf{z}[n]$ and the Riccati equation (2.70). Defining the *a priori estimation error* $\xi[n]$[6] by

$$\xi[n] = y[n] - \hat{\mathbf{h}}[n-1]^{T}\,\mathbf{x}[n] \tag{2.71}$$

the recursive update equation for the tap-weight vector can be written as:

$$\hat{\mathbf{h}}[n] = \hat{\mathbf{h}}[n-1] + \mathbf{k}[n]\,\xi[n] \tag{2.72}$$

For the RLS to be applicable, the start value of the inverse correlation matrix $\mathbf{P}[0]$ must assure the nonsingularity of the correlation matrix $\mathbf{\Phi}[n]$. Therefore $\mathbf{\Phi}[0]$ will be set to

$$\mathbf{\Phi}[0] = \delta^{-1}\,\mathbf{I} \tag{2.73}$$

where $\mathbf{I}$ is the $(L \times L)$ identity matrix and δ is a small positive constant, with recommended value $\delta \ll 0.01\sigma_x^2$, where σ_x^2 is the variance of a data sample $x[n]$ [Haykin 96]. For the initial value of the tap-weight vector usually

$$\hat{\mathbf{h}}[0] = \mathbf{0} \tag{2.74}$$

will be taken, with $\mathbf{0}$ being the $(L \times 1)$ null vector.

The RLS algorithm may be summarized as follows [Vaseghi 96]:
Input signals: $\mathbf{x}[n], y[n]$

[6] The *a priori estimation error* is different from the *a posteriori estimation error* $e[n]$ defined by $e[n] = y[n] - \hat{\mathbf{h}}[n]^{T}\,\mathbf{x}[n]$.

Tap-weights: $\hat{\mathbf{h}}[n]$
Initialization:

$$\mathbf{\Phi}[0] = \delta\mathbf{I} \quad (2.75)$$

$$\hat{\mathbf{h}}[0] = \mathbf{0} \quad (2.76)$$

Computation for each instant of time $n = 1, 2, \ldots$

1. *filter gain vector:*

$$\mathbf{k}[n] = \frac{\lambda^{-1}\mathbf{P}[n-1]\,\mathbf{x}[n]}{1+\lambda^{-1}\mathbf{x}^T[n]\,\mathbf{P}[n-1]\,\mathbf{x}[n]} \quad (2.77)$$

2. *estimation error:*

$$\xi[n] = y[n] - \hat{\mathbf{h}}[n-1]^T\,\mathbf{x}[n] \quad (2.78)$$

3. *tap-weight adaptation:*

$$\hat{\mathbf{h}}[n] = \hat{\mathbf{h}}[n-1] + \mathbf{k}[n]\,\xi[n] \quad (2.79)$$

4. *inverse correlation matrix update:*

$$\mathbf{P}[n] = \lambda^{-1}\,\mathbf{P}[n-1] - \lambda^{-1}\,\mathbf{k}[n]\,\mathbf{x}^T[n]\,\mathbf{P}[n-1] \quad (2.80)$$

Convergence Rate

The RLS algorithm converges exponentially and uniformly, regardless of the eigenvalue spread, i.e. the condition number of the ensemble-averaged correlation matrix $\mathbf{R}$ of the input signal $\mathbf{x}[n]$. After a change in the input, convergence depends only on the weighting factor λ [Marple 87].

The ensemble-averaged learning curve of the RLS algorithm converges in about $2L$ iterations [Haykin 96], where L is the number of taps in the transversal filter. This means, that the RLS algorithm converges an order of magnitude faster than the LMS algorithm.

In high noise, or where the condition number is low, this improvement in the rate of convergence is not achieved. In this case the LMS and RLS algorithms have comparable convergence rates.

Tracking Capability

In a nonstationary environment when the RLS algorithm has to *track* the statistical variations of the input data, the performance of the algorithm depends very much on the mismatch between the considered mathematical model and the physical process which generates the input data.

To enhance the tracking capability of the RLS algorithm, i.e. to minimize the mismatch, every supplementary knowledge of the input data generating process should be considered in the model definition.

Computational Load

Due to the matrix updates in Eqs. (2.69) and (2.70) the RLS algorithm requires a number of computations (multiplies/adds) per time update proportional to L^2, where L is the number of adjustable weights in the algorithm. Compared to the computational load of the LMS of $2L$ this is an extremely high computational effort.

Therefore *fast RLS* algorithms have been developed, that reduce the computational power from L^2 to $8L$ operations per time update. This reduction in complexity is achieved by considering the redundancy in the Toeplitz structure of the input data matrix and by exploiting this redundancy through the use of an additional backward linear prediction. A drawback of the fast RLS algorithm is its poor long-term numerical stability. The fast RLS is accurate only for short- to medium-length data [Marple 87]. When implemented in fixed point arithmetic, the unpredictable round-off errors accumulate until they destroy the proper operation of the algorithm. In [Schütze & Ren 92] a detailed comparison of eight known fast recursive least squares algorithms is presented and the different numerical sensitivity of the algorithms is demonstrated.

Chapter 3

Speech Enhancement

One of the main problems with mobile handsfree operation in a car is the high background noise level which degrades the system perfomance.
Conversational speech has an average of 60 dB *Sound Pressure Level (SPL)* and in a vehicle it is subject to a variety of corruption processes, such as other acoustic sources with spectral content overlapping that of speech, convolution with time-varying transfer function paths or modulation due to reflection from vibrating surfaces [Campbell 93]. At higher cruising speeds or during hard accelerations, the background noise level is often around 70 dB or even higher [Häkkinen & Väänänen 93]. In such cases, the speech signal to background noise ratio at the microphone is very low or even negative. Usually the speech is still intelligible, but the amount of background noise is annoying to the far-end listener.

In a car, the noise model can be considered as a combination of several independent noise sources [Lockwood et al. 91] caused by engine and tyres with mostly low frequency components and aerodynamic turbulences with a broader spectrum. In the telephone audio frequency band there are no clearly dominant noise sources, the noise field is diffuse. The experiments reported in [Goubran et al. 90] indicate, that low frequency noise signals are highly correlated and that the correlation decreases with increasing frequency until it vanishes for frequencies higher than about 2 kHz. It was also found that there is a significant correlation between the acoustic noise in the area facing the driver's seat and the noise in other locations of the car.
The graphs in Figure 3.1 and 3.2 show time and frequency domain representations of noise recorded in vehicles.

Speech enhancement algorithms may improve speech perception by improving speech quality, increasing the intelligibility and/or reducing listener fatigue [Munday 88]. Noise reduction is primarily intended to achieve an increase in intelligibility and a reduced listener fatigue. Subjectively, the quality of the en-

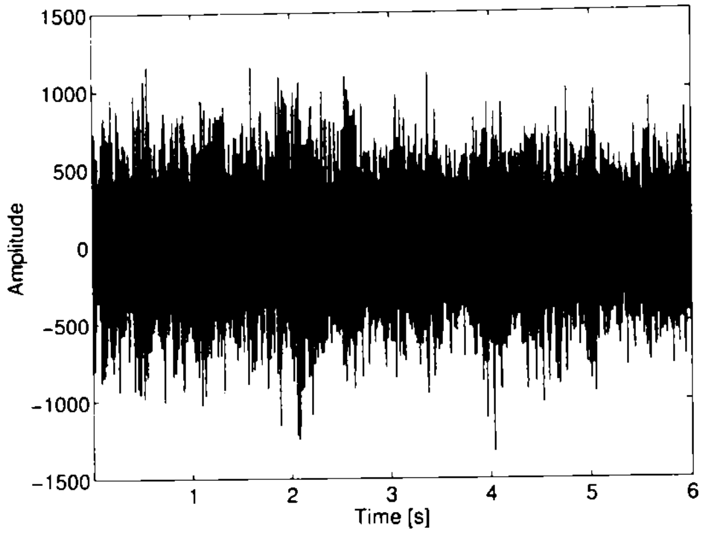


Figure 3.1: Car noise, high way, in the time domain

hanced near-end audio signal depends both on the attenuation and sound of the background noise. Ideally, the sound of the background noise should not change by the attenuation. If changes occur, they should not make the residual noise sound like human voice or musical tones, or anything disturbing. The noise attenuation or removal may also distort the speech, which the user will not tolerate, if it is clearly noticeable. A known person's voice should not sound odd, heard through the noise attenuation system. Thus, the main objective when designing speech enhancement systems is to maximize noise reduction, while keeping the introduced distorsions at an acceptable, not annoying level. The naturalness of the residual noise is also very important.

In [O'Shaughnessy 89], speech enhancement methods are considered to fit into three general classes, each with its own advantages and limitations:

- subtraction of interfering sounds
- suppression of nonharmonic frequencies
- resynthesis using vocoders

If an interfering noise can also be captured separately from the desired speech, the latter is usually enhanced by subtracting out the former. Noise subtraction usually requires a second microphone, which is placed closer to the noise source than the primary microphone recording the desired speech. This second recording provides the noise reference, which may be subtracted from the primary recording after processing. If only a single recording is available, analysis during speech pauses can furnish an estimate of the noise.

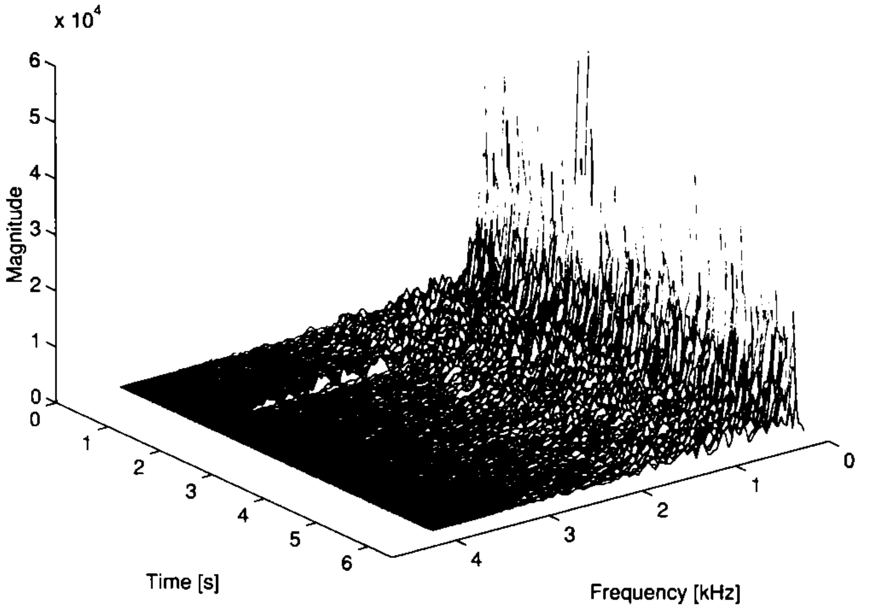


Figure 3.2: Car noise from Figure 3.1 in the spectrogram representation

The harmonic enhancement method attempts to identify the fundamental frequency F_0 of either the desired speech or an interfering source. If the desired sound is the strongest component in the signal, its frequencies can be identified and other frequencies may then be suppressed. If an interfering sound is the strongest, then its frequencies can be identified and suppressed, considering that the remaining frequencies presumably contain the desired speech.

The resynthesis using vocoders cleans speech signals by parametric estimation and speech resynthesis. Speech synthesizers generate noise-free speech from parametric representations of either vocal tract model or previously analyzed speech. Standard methods such as linear predictive coding do not replicate the spectral envelope precisely, but usually preserve enough information to yield good output speech. Such synthesis often has a mechanical sound quality, but is free of interference and quite intelligible, presuming the parameters represent the original speech accurately.

When classifying the speech enhancement methods considering the number of microphones used, there are

- systems with a single microphone
- systems with two or more microphones.

Speech enhancement systems with a single microphone have the advantage of us-

ing simple standard recording equipment, but which can cancel only stationary noise and which require a positive SNR [Van Compernolle 92] of the input signal. For these systems a speech/noise or *Voice Activity Detector (VAD)* is indispensable, but the segmentation process at low SNR is very difficult and the detector can conceivably take the wrong decision.
Another drawback is that the noise estimated during speech pauses is not re-estimated during speech, since the method assumes that noise during speech pauses is representative of noise during periods of speech. Thus, rapidly varying noise can cause problems.
The ”one microphone approach” gives much less enhancement than the two microphone method, because it employs an average spectral model of the noise and is able to identify only the spectral distribution of noise energy, not its time variation.
Systems with two or more microphones need more hardware and some knowledge about the place of the desired source, but permit cancellation of also nonstationary or very strong interfering noise.

The speech enhancement can be implemented either in the time domain or in the frequency domain. As human understanding is much better in the spectral domain than in the time domain [Van Compernolle 92] and because of the short-time stationarity of speech, modifications to the signal are best performed in the spectral domain. Due to psychoacoustic properties of the ear, the noisy phase may be used for transforming the signal back into the time domain. Phase is unimportant in speech enhancement as long as local signal-to-noise ratios are at least about 6 dB [Vary 85].

3.1 Single Microphone Noise Suppression

In applications where only the noisy signal is available, the random noise cannot be cancelled out, but using the statistics of the noise, it is possible to reduce the average effects of the noise on the signal spectrum [Vaseghi 96]. The effect of additive noise on the magnitude spectrum of a signal consists in an increase of the mean and of the variance of the spectrum.
The increase in the mean of the noisy signal spectrum can be reduced by subtracting an estimate of the mean of the noise spectrum.
The increase in variance of the noisy signal spectrum, due to the random fluctuations of the noise, cannot be cancelled out. The noise variations can be reduced by averaging the noisy signal frequency components, but this will lead to a reduction in the time resolution of the nonstationary spectral events. As time resolution plays a very important part in both quality and intelligibility of audio signals, the averaging process should reflect a compromise between these two conflicting requirements.

Considering the noise being additive and stationary, the model of the noisy signal in the time domain is given by

$$y[n] = x[n] + n[n] \tag{3.1}$$

where $y[n]$, $x[n]$ and $n[n]$ are the noisy signal, the original signal and the additive noise respectively, and n is the discrete time index.
In the frequency domain the model will be

$$Y[k] = X[k] + N[k] \tag{3.2}$$

with $Y[k]$, $X[k]$ and $N[k]$ representing the Fourier transforms of the noisy signal, the original signal and the noise respectively, or, in other words, the short-time spectra associated with the windowed signals $y[n]$, $x[n]$ and $n[n]$. k denotes the frequency bin number.

Investigations on several different single input processing techniques, performed in the frequency domain are presented in [Curtis & Niederjohn 78]. An important conclusion of this study is that any weighting function used on the frequency magnitude spectrum must be relatively smooth between adjacent spectral lines to avoid the introduction of distorsions. Placing limits of 0.6 to 1.66 on the permissible change between successive points in the processed noisy speech spectrum leads to satisfaying results.
Another important issue when processing noisy speech in the frequency domain is the choice of the length and type of the windowing function applied to the time domain noisy signal before the Fourier transform.
The duration of the windowed speech segment must be short enough, so that speech can be considered stationary. On the other hand, the random variations in amplitude of the noise component increase as the segment duration is shortened. These variations from frame to frame and from bin to bin are considerable even though the mean noise spectrum may be stationary and smooth [Xydeas et al. 88].
Windowing the noisy speech signal in time domain with a rectangular function can distort the signal in an unacceptable way. This is because the transformed window is a sinc function with high sidelobes and its convolution with the Fourier transform of the noisy signal generates relatively large erroneous spectral lines in the neighbourhood of the true components [Munday 88]. This effect is called frequency leakage and leads to a severe limitation of the dynamic range of the true components. If windowing functions with superior leakage performance are used, then the disadvantage of reduced spectral resolution must be overcome. This can be equalized by overlapping the time domain blocks before applying the Fourier transform. The equalization procedure is completed after inverse transformation, when the overlap-add operation is performed. This procedure leads to an increased processing burden, since each time domain sample is used twice.

Thus, a compromise must be found in the choice of the windowing function between frequency leakage and spectral resolution.

3.1.1 Spectral Subtraction

Because of its ease of implementation and relatively good performance, spectral subtraction is one of the most widely employed speech enhancement techniques. In spectral subtraction the incoming signal $x[n]$ is buffered and divided into segments of N samples length. In order to alleviate the effects of the discontinuities at the endpoints, each segment is windowed. The windowed segments are then transformed, via *Discrete Fourier Transform (DFT)* to N spectral samples. The window length is chosen to be about as long as an average speech segment, i.e. 20 - 30 ms.
The approach is based on subtracting the magnitude or power spectrum of a noise-only record, or an average of records, from that of a noisy speech record [Lim & Oppenheim 79]. The result is combined with the phase of the original noisy speech and inverse transformed to get the enhanced signal. Some residual noise is generated, because no accurate noise information is available. Due to changes in the noise spectrum, negative values can occur after subtraction which may be set to zero or some small value. The resulting noise spectrum then contains randomly appearing spectral lines which generate short tone bursts resulting in the disturbing artifact known as *musical noise.*

The equation describing the spectral subtraction is [Berouti et al. 79]

$$|\hat{X}[k]|^b = |Y[k]|^b - \alpha\,\overline{|N[k]|^b} \tag{3.3}$$

where $|\hat{X}[k]|^b$ is an estimate of the original signal spectrum and $\overline{|N[k]|^b}$ is the time averaged noise spectrum. Depending on the value of b there can be defined

- the magnitude spectral subtraction, for $b = 1$ [Boll 79]
- the power spectral subtraction, for $b = 2$.

Thus, spectral subtraction may be performed in the magnitude or power spectral domains. The difference between the two implementations is rather small, but according to [Berouti et al. 79], the spectral subtraction with $b = 2$ was found to yield better output quality, in general.
The *subtraction parameter* $\alpha \geq 1$ in Eq. (3.3) controls the amount of noise subtracted from the noisy signal. For $\alpha = 1$ a full subtraction, for $\alpha > 1$ an oversubtraction are performed.

The most difficult task in spectral subtraction is the extraction of a good noise power spectrum estimate out of the noisy speech signal. This is done in a two-step process [Berouti et al. 79]:

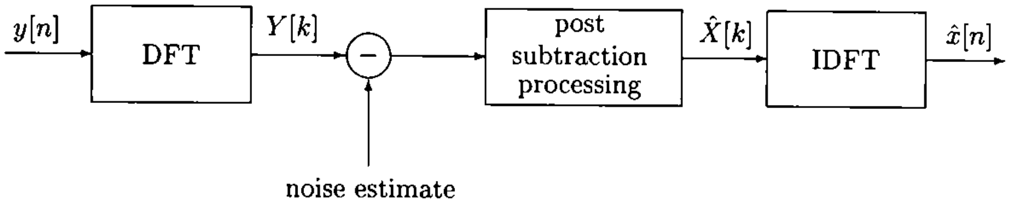


Figure 3.3: Block diagram of spectral subtraction algorithm

1. averaging short-time power spectra over frames that are supposed to contain only noise. The time-averaged noise spectrum can be expressed as follows:

$$\overline{|N[k]|^b} = \frac{1}{M}\sum_{i=0}^{M-1} |N_i[k]|^b \tag{3.4}$$

   where $|N_i[k]|^b$ represents the spectrum of the i-th noise frame, and it is assumed that there are M frames in a noise-only period. M is variable from one noise period to the next.

2. the noise spectrum can be estimated at each frequency bin by a first order digital lowpass filter with a suitable time constant:

$$\overline{|N_i[k]|^b} = \rho\,\overline{|N_{i-1}[k]|^b} + (1-\rho)\,|N_i[k]|^b \tag{3.5}$$

   where i refers to the frame number and ρ is the lowpass filter coefficient, typically set between 0.8 and 0.95 [Vaseghi 96].

To be sure that a given frame contains no speech, an adequate noise/speech discriminator must be used. The simplest approach to noise detection is to compare the signal frame energy to a threshold. This threshold should be adaptive, because background noise level changes with time. The detector's decision should be conservative, in that it is less harmful to classify noise as speech then the other way round.

As already mentioned, spectral subtraction may produce negative estimates of the magnitude or power spectrum. This is more probable as the signal-to-noise ratio (SNR) decreases. In order to reduce the spectral excursions, the post-processing block of Figure 3.3 performs a mapping function $T[\cdot]$ of the form:

$$T\left[|\hat{X}[k]|^b\right] = \begin{cases} |\hat{X}[k]|^b & \text{if} \quad |\hat{X}[k]|^b > \beta\,\overline{|N[k]|^b} \\ fn\left[\overline{|N[k]|^b}\right] & \text{otherwise} \end{cases} \tag{3.6}$$

In its simplest form the noise-dependent function will be [Berouti et al. 79]:

$$fn\left[\overline{|N[k]|^b}\right] = \beta\,\overline{|N[k]|^b} = \text{noise floor} \tag{3.7}$$

where $0 < \beta < 1$ is called the *spectral floor parameter*.
Overestimation of the noise and application of a non-zero threshold level allows some manipulation of the musical noise. Doing this will trade the disturbing distorsions for a somewhat less annoying noise.
After performing the spectral domain modifications, the magnitude spectrum estimate $|\hat{X}[k]|$ is combined with the phase of the noisy signal and transformed back into the time domain via an *Inverse Discrete Fourier Transform (IDFT)*.

Reducing Processing Distorsions

The main problem in spectral subtraction [Lockwood et al. 91] is the presence of musical tones, introduced both by the analysis based on the short-time spectral representation of the signals involved and the nonstationarity of the noise. As no reevaluation during speech periods is done for the noise estimate computed during speech pauses, this estimate will be no longer valid if the noise varies rapidly. The dominant distorsion is mainly due to the nonlinear mapping of the negative or small valued spectral estimates. Analyzing the characteristics of audio signals such as speech or music, significant differences between the authentic audio signals and the annoying musical noise may be observed [Vaseghi 96]. Musical notes tend to be short lived, random isolated bursts of narrow band signals with relatively small amplitudes. These differences may be used to identify and remove the musical tones. A well-known technique consists in examining the frequency components in time. If a frequency component has a duration shorter than a preselected time window, an amplitude smaller than a threshold, then it is considered to be a distorsion and removed.

The distorsions due to the variations of the noise spectrum can be reduced by lowpass-filtering the magnitude spectrum of the noisy speech at each frequency bin. As every averaging process has the undesirable effect of smearing and obscuring the time variations of the signal spectrum, a compromise has to be found between the conflicting requirements of reducing the noise variance and of retaining the time resolution of the nonstationary signal.

Nonlinear Spectral Subtraction

Another approach to the reduction of musical tones associated with the spectral subtraction method is to use in Eq. (3.3) a signal-to-noise ratio (SNR)-dependent subtraction factor α. It has been observed that at low SNR, when the signal may be considered as lost in noise, oversubtraction ($\alpha > 1$) followed by a nonlinear implementation of the subtraction process can produce improved results [Lockwood & Boudy 92].
The nonlinear variant of spectral subtraction can thus be expressed by the fol-

lowing equation:

$$|\hat{X}[k]|^b = |Y[k]|^b - \alpha(SNR[k])\,\overline{|N[k]|^b} \tag{3.8}$$

where $\alpha(SNR[k])$ represents the SNR-dependent subtraction factor.
There are known several forms of the SNR-dependent subtraction factor discussed in the literature, the most important being the following:

- the subtraction factor $\alpha(SNR[k])$ depends on the mean and the variance of the estimated noise

  $$\alpha(SNR[k]) = \left(1 + \frac{sd(|N[k]|)}{\overline{|N[k]|}}\right) \tag{3.9}$$

  In this case, the amount oversubtracted is the standard deviation of the noise [Vaseghi 96].

- the method presented in [Lockwood & Boudy 92], which considers the term

  $$\overline{|N[k]|^b}_{NL} = \alpha(SNR[k])\,\overline{|N[k]|^b} \tag{3.10}$$

  as a function of the maximum value of noise spectrum over M frames and the signal-to-noise ratio:

  $$\begin{aligned}\overline{|N[k]|^b}_{NL} &= \Phi\left(\max_{over\,M\,frames}(|N[k]|^b),\ SNR[k],\ \overline{|N[k]|^b}\right)\\ &= \frac{\max_{over\,M\,frames}(|N[k]|^b)}{1+\gamma\ SNR[k]}\end{aligned} \tag{3.11}$$

  where γ is a design parameter.
  $\Phi(\cdot)$ is a nonlinear function weighting the subtraction process according to the signal to noise ratio of a specific frequency bin. For decreasing SNR, the output of the nonlinear estimator $\Phi(\cdot)$ approaches $\max_{over\,M\,frames}(|N[k]|^b)$, while in high SNR environment it approaches zero.
  For oversubtraction, $\Phi(\cdot)$ is limited to

  $$\overline{|N[k]|^b} \leq \Phi\left(\max_{over\,M\,frames}(|N[k]|^b),\ SNR[k],\ \overline{|N[k]|^b}\right) \leq 3\,\overline{|N[k]|^b} \tag{3.12}$$

  Any arbitrary function, implementing the idea of applying a minimum subtraction factor in high SNR regions and subtracting more noise in regions with low SNR, can be chosen as $\Phi(\cdot)$.
  The spectral floor, an important factor that will prevent the subtraction result of becoming negative, will be implemented as follows:

  $$|\hat{X}[k]|^b = \begin{cases} |\hat{X}[k]|^b & \text{if} \quad |\hat{X}[k]|^b \geq \beta\,\overline{|N[k]|^b} \\ \beta\,\overline{|N[k]|^b} & \text{otherwise} \end{cases} \tag{3.13}$$

  A typical value for β is 0.1.

3.1.2 Wiener Filtering

This approach to speech enhancement is based on the use of Wiener filter, which, when dealing with stochastic wide-sense stationary signals, provides a *least mean square error (LMSE)* estimate of the desired signal. In this sense, the degraded speech is used to obtain a filter which is then applied either in the time domain or in the frequency domain to get an estimate of the undegraded speech [Lim & Oppenheim 79].
The estimate $\hat{X}[k]$ of the short-time spectrum of the original speech will take the form

$$\hat{X}[k] = W[k]Y[k] \tag{3.14}$$

where $W[k]$, the noncausal Wiener filter is approximated with the adaptive Wiener filter with the frequency response

$$W[k] = \frac{E[|X[k]|^2]}{E[|X[k]|^2] + E[|N[k]|^2]} \tag{3.15}$$

Making use of a short-time estimate of the measurement power spectrum and a long-term estimate of the noise spectrum [Van Compernolle 92], $E[|X[k]|^2]$ can be estimated. The frequency response can then be rewritten as

$$W[k] = \frac{E[|Y[k]|^2] - E[|N[k]|^2]}{E[|Y[k]|^2]} \tag{3.16}$$

Comparing the Wiener filter to the spectral subtraction filter defined as

$$H[k] = \frac{|Y[k]|^b - \overline{|N[k]|^b}}{|Y[k]|^b} \tag{3.17}$$

it can be seen, that the Wiener filter is based on the *ensemble average* spectra of the signal and noise, while the spectral subtraction filter uses the instantaneous spectra of the noisy signal and the *time-averaged* spectra of the noise. For an ergodic process the spectral subtraction filter approaches the Wiener filter.
As in practice the signals are nonstationary, the averaging nature of the mean square error criterion is not well suited. High coefficient update rates generate musical noise artifacts, while low rates result in perceived convolutional distortion of the speech [Campbell 93].
Starting from Eq. (3.15) it can be stated that for additive noise the Wiener filter attenuates each frequency component in proportion to an estimate of the signal to noise ratio $SNR[k]$

$$W[k] = \frac{SNR[k]}{SNR[k] + 1} \tag{3.18}$$

This means that for a noise-free signal (high SNR) the attenuation is small or inexistent, i.e. $W[k] \approx 1$, and for an extremely noisy signal $W[k]$ will tend to zero. Hence, for additive white noise, the Wiener filter response will approximately follow the signal spectrum [Vaseghi 96].

3.1.3 Spectral Scaling

Spectral scaling consists in applying a non-linear transfer function to each component of the short-time spectrum of the noisy speech in order to reduce the contribution of those spectral components likely to be noise dominated [Munday 88]. The use of a suitably positioned threshold on the spectral magnitudes permits the cancellation of additive noise. This thresholding is equivalent to applying a non-linear input-output transfer characteristic, which passes all spectral components above the specified value and attenuates those below the specified threshold value.

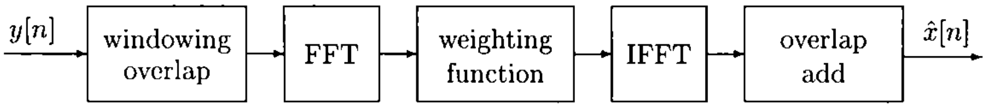


Figure 3.4: Block diagram of spectral scaling algorithm

The noisy speech signal is segmented into consecutive frames which may also be overlapped, and then a Fast Fourier Transform (FFT) is performed on each frame. Each of the spectral lines (bins) of the short-term spectrum of the noisy signal is modified according to a particular weighting law, with the aim of attenuating the energy of the noise while preserving the desired speech signal. Signal reconstruction in the time domain is performed by computing an inverse FFT on the processed spectrum (Figure 3.4).

The modification to the spectral lines is dependent on the magnitude of the frequency bin and on the noise magnitude at that frequency, estimated from the noisy input signal during periods when no speech activity is detected. In other words, the attenuation depends on the signal to noise ratio of the respective frequency bin.
A large attenuation of noise can be achieved by employing a non-linear weighting function. It will leave magnitude samples unchanged that are large compared to the noise estimate and will heavily attenuate small samples which are considered to be mainly due to the noise. The weighting function will show two regions, a non-linear and a linear region. The non-linear region introduces a smoothly changing attenuation, the lower the magnitude of the frequency bin, the greater the degree of attenuation. The linear region is applied to the larger magnitude samples.
A significant noise reduction can only be achieved at the expence of a quality deterioration of the speech signal: the better the noise elimination, the more distorted the reconstructed speech. A balance must be found between noise reduction and speech distorsion, so that the intelligibility of speech is maximized.

In [Crozier et al. 93] the following weighting function is presented:

$$|\hat{X}[k]| = (|Y[k]|)^{1-\gamma} \left(\chi \, |\hat{N}[k]|\right)^{\gamma} \tag{3.19}$$

where $|\hat{X}[k]|$, $|Y[k]|$ and $|\hat{N}[k]|$ are the short-term magnitude spectra of the enhanced speech, the noisy speech and the estimated noise respectively. γ is defined as

$$\gamma = \begin{cases} 0 & \text{if} \quad |Y[k]| \geq \chi \, |\hat{N}[k]| \\ \kappa \left(1 - \dfrac{|Y[k]|}{\chi \, |\hat{N}[k]|}\right) & \text{otherwise} \end{cases} \tag{3.20}$$

The *scaling factor* χ specifies how large an input magnitude must be before it is left unchanged by the weighting function. The larger χ the better the noise reduction, but the more distorted the speech signal will be.
The *exponential constant* κ controls the degree of attenuation of frequency bins with a magnitude less than $\chi \, |\hat{N}[k]|$. γ varies linearly from 0 to κ.

It has been shown [Xydeas et al. 88] that, when applying spectral scaling to the different frequency bins, it is useful to also consider the spectra of adjacent frames. This can considerably reduce the variation of those bins which are mostly due to the background noise, thus allowing a better attenuation of noise energy for a given level of speech distorsion.

3.1.4 Linear Prediction

The basic idea behind linear prediction is that a sample $x[n]$ of the signal $x[i]$, with $i = 1, 2, \ldots, N$ can be forecasted at time n using a linearly weighted combination of P past samples $x[n-1], x[n-2], \ldots, x[n-P]$ as

$$\hat{x}[n] = \sum_{k=1}^{P} a_k \, x[n-k] \tag{3.21}$$

where n is the discrete time index, $\hat{x}[n]$ is the prediction of $x[n]$ and a_k are the prediction coefficients. The difference between the actual sample $x[n]$ and the predicted sample $\hat{x}[n]$ is called the *residual* or *prediction error* $e[n]$ and is given by

$$e[n] = x[n] - \sum_{k=1}^{P} a_k \, x[n-k] \tag{3.22}$$

By minimizing the sum of the squared error signals over a finite interval, a unique set of predictor coefficients can be determined [Rabiner & Schafer 78].

Linear prediction is very closely related to the basic speech production model.

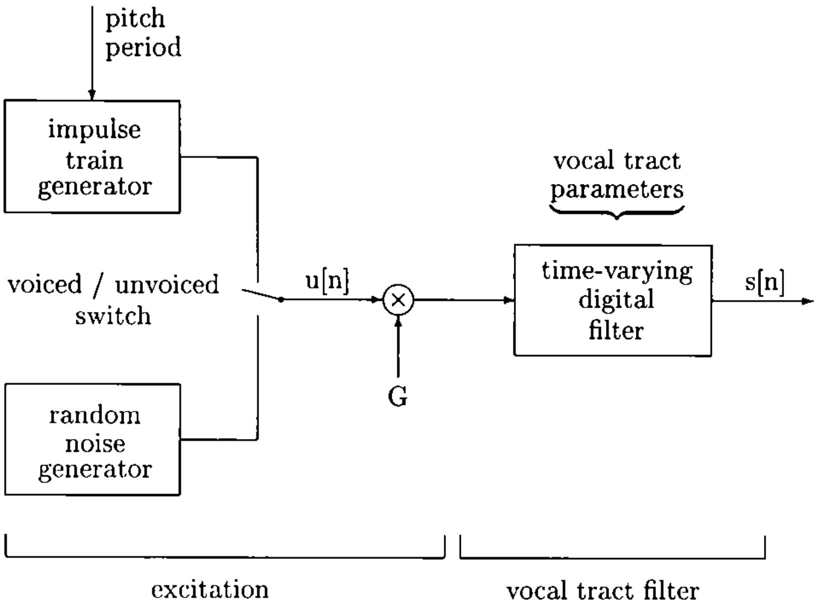


Figure 3.5: Block diagram of simplified model for speech production

Corresponding to this model (Figure 3.5), speech can be considered as the output of a linear, time-varying system excited by either quasi-periodic pulses during periods of voiced speech, or random noise during unvoiced speech. Linear prediction provides a robust method for estimating the parameters that characterize the linear, time-varying system. Once the predictor coefficients have been determined, the vocal tract system has been uniquely identified, so that it can be modelled as an all-pole linear system with the transfer function [Lim & Oppenheim 78]:

$$H[z] = \frac{S[z]}{U[z]} = \frac{G}{1 - \sum_{k=1}^{P} a_k \, z^{-k}} \tag{3.23}$$

with G being the gain parameter and a_k the coefficients of the digital filter. These parameters are slowly varying with time.

For a sufficient high order of P, the all-pole model provides a good representation for almost all the sounds of speech, nasals and fricatives included, which according to acoustic theory would require both zeros and poles in the transfer function of the vocal tract.

Thus, the speech waveform $s[n]$ is assumed to satisfy a difference equation of the form

$$s[n] = \sum_{k=1}^{P} a_k \, s[n-k] + G \, u[n] \tag{3.24}$$

where $u[n]$ is the input excitation of the system.

The all-pole predictor model (3.23) transforms an uncorrelated excitation signal, $u[n]$, into a correlated signal $s[n]$ [Vaseghi 96].

The inverse linear predictor, as the name implies, transforms a correlated signal $x[n]$ back into an uncorrelated signal $e[n]$

$$e[n] = x[n] - \hat{x}[n]. \tag{3.25}$$

The inverse filter, also known as the prediction error filter, is an all-zero, finite impulse response filter with the following transfer function:

$$A[z] = 1 \ - \ \sum_{k=1}^{P} a_k \, z^{-k} \tag{3.26}$$

The linear prediction filter $H[z]$ and the inverse filter $A[z]$ are thus related as follows:

$$H[z] = \frac{G}{A[z]} \tag{3.27}$$

The inverse filter of an all-pole filter, being an all-zero filter with the zeros situated at the same angular frequencies as the poles of the all-pole filter, has the effect of flattening the spectrum of the input signal. It is therefore also known as a spectral whitening or decorrelation filter.
The basic problem of linear prediction analysis is to define a set of prediction coefficients a_k directly from the speech signal in order to get a good estimate of the speech properties. The predictor coefficients will minimize the mean-squared prediction error over short segments of speech for which stationarity can be assumed [Rabiner & Schafer 78]. The equations thereby obtained are

$$\sum_{k=1}^{P} a_k \sum_{n} s[n-i] \, s[n-k] = \sum_{n} s[n-i] \, s[n] \tag{3.28}$$

for $i = 1, 2, \ldots, P$. The range of summation is a finite interval that depends on the method used in solving this set of P equations in P unknowns.
Defining

$$\phi[i,k] = \sum_{n} s[n-i] \, s[n-k] \tag{3.29}$$

equation (3.28) can be written more compactly as

$$\sum_{k=1}^{P} a_k \, \phi[i,k] = \phi[i,0] \qquad i = 1, 2, \ldots, P \tag{3.30}$$

The most important procedures for solving the linear prediction analysis Eqs. (3.30) are

- the *Cholesky* decomposition for the covariance method
- *Durbin's* recursive solution for the autocorrelation method

- *Burg's* procedure for the lattice method.

In [Rowden & Hall 91] and [Rabiner & Schafer 78] comparisons between the procedures of solving the linear prediction analysis equations are presented. Computational considerations, numerical and physical stability of the solutions are also compared. Each method has its own advantages and limitations.
Despite the benefit of not requiring a window function, the covariance method is not used much in speech analysis. The reason is the unstable filter configurations that can be produced under certain circumstances.
The autocorrelation method requires a longer data frame because of windowing and overlapping, but it produces stable filters. This is, however, at a cost of some loss of clarity of the speech due to the windowing process.
The lattice method produces stable, clear speech, but is computationally expensive.
For quasi-stationary signals, such as voiced speech, two types of correlations can be considered which allow a more accurate prediction. These are

- the short-term prediction, which uses the correlation of each sample $x[n]$ with the P immediate past samples $x[n-1], \ldots, x[n-P]$

- the long-term prediction, which is the correlation of a sample $x[n]$ with e.g. $2Q+1$ similar samples $x[n-T+Q], \ldots, x[n-T-Q]$ a pitch period T away [Vaseghi 96]. The long-term correlation may be modelled by a pitch predictor defined as

$$\hat{x}[n] = \sum_{k=-Q}^{Q} p_k \, x[n-T-k] \tag{3.31}$$

  where p_k are the coefficients of a long-term predictor of order $2Q+1$. The pitch period T can be obtained from the correlation function of $x[n]$, it is the first nonzero time lag where the correlation function attains a maximum.

Combining the short-term and long-term predictors into a single model, $x[n]$ can be written as

$$x[n] = \underbrace{\sum_{k=1}^{P} a_k \, x[n-k]}_{short-term\ prediction} + \underbrace{\sum_{k=-Q}^{Q} p_k \, x[n-k-T]}_{long-term\ prediction} + \varepsilon[n] \tag{3.32}$$

with $\varepsilon[n]$ representing the prediction error of the long term filter. In this model, each sample is expressed as a linear combination of P immediate past samples and $2Q+1$ samples a pitch period away.

Linear prediction is mostly used in speech coding, but it can also provide some means of enhancing speech corrupted by additive noise.

For a noisy signal, linear prediction analysis models the combined spectra of the signal and the noise. The estimated coefficients can be used in the restoration of a signal observed in additive noise.
An illustration of an iterative implementation of a signal restoration system based on a linear prediction model of speech [Vaseghi 96] is presented in Figure 3.6.
The iteration process starts with the estimation of the predictor parameters $\hat{a}_k$

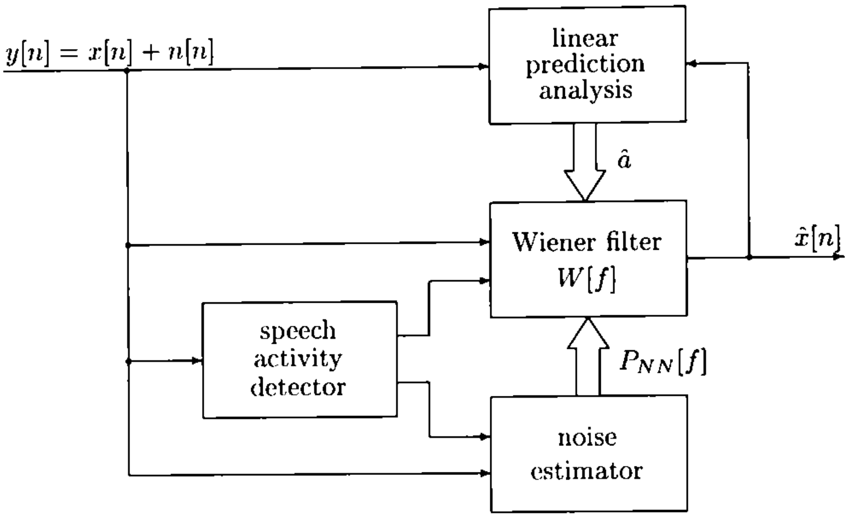


Figure 3.6: Signal restoration system based on a linear prediction model of speech

from the noisy speech. For determining the gain G of the linear prediction model, Parseval's theorem[1] is used

$$\frac{1}{N}\sum_{f=0}^{N-1}\frac{\hat{G}^2}{\left|1-\sum_{k=1}^{P}\hat{a}_{k,i}\,e^{-j2\pi kf/N}\right|^2}=\sum_{n=0}^{N-1}y^2[m]-\hat{E}_n \tag{3.33}$$

where $\hat{a}_{k,i}$ are the coefficient estimates at iteration i and $\hat{E}_n$ is the estimate of noise energy determined during periods when no speech is present in the noisy signal. Having the prediction coefficients and the gain of the linear prediction model, an estimate of the power spectrum of the speech model $\hat{P}_{X_iX_i}[f]$ at frequency bin f can be calculated.
The estimation of the noise power spectrum can be performed during periods when no speech is present. Having the estimates of the power spectrum of speech and the estimate of the noise power spectrum, the Wiener filter frequency response can be calculated:

$$\hat{W}_i[f]=\frac{\hat{P}_{X_iX_i}[f]}{\hat{P}_{X_iX_i}[f]+\hat{P}_{N_iN_i}[f]} \tag{3.34}$$

[1] Parseval's theorem allows to equate the total power or energy of a signal in the time and frequency domains [Lynn & Fuerst 89].

where i is the iteration step and f represents the frequency bin number. The magnitude spectrum of the estimated noise-free signal will then be

$$\hat{X}_{i+1}[f] = \hat{W}_i[f]\,Y[f] \tag{3.35}$$

Combining $\hat{X}_{i+1}$ with the phase of the noisy signal, the time domain signal for iteration step i will be restored. This procedure will be repeated until convergence, or for a specified number of iterations.

Linear prediction analysis also may be applied to noisy signals in combination with other speech enhancement techniques, e.g. for reducing the musical noise after spectral shaping [Crozier et al. 93]. Most of the energy in a segment of voiced speech is contained within the formants. Musical tones in these regions will be masked out by the high energy speech harmonics, so the majority of disturbing musical noise will be in the regions between the formants. The described enhancement method uses a weighting function derived from the estimated formant distribution given by the linear prediction spectrum of speech. This weighting function is used to further attenuate spectral regions with low speech energy. For sentences corrupted by high ambient noise (SNR of 0 dB), the proposed algorithm cannot reduce the musical tones because the linear prediction approximation becomes inaccurate.

Another approach to the enhancement of a noise degraded signal by using linear prediction is described in [Richardson & Gowdy 96]. The output of the proposed enhancement system is a linear combination of the actual speech input and a synthesized speech signal. The latter is generated using the *Linear prediction coefficients (LPC)* and an estimate of the excitation from the current speech. The algorithm improves speech quality by emphasizing the speech signal rather than removing the unwanted noise. The most difficult problem is the definition of an excitation filter which reconstructs the higher harmonics of the glottal excitation for voiced phonemes[2].

3.2 Multimicrophone Noise Cancellation

3.2.1 Adaptive Noise Cancellation

The method of adaptive noise cancellation [Widrow et al. 75] makes use of a *primary* input containing the corrupted speech signal and one or more *auxiliary* or *reference* inputs containing only noise [Goubran et al. 90], correlated in some way with the noise from the primary input. The reference input is adaptively filtered and subtracted from the primary signal to obtain the signal estimate. When

[2] See Annex A.

suitable input signals are available the method allows enhancement of speech degraded by additive noise or interference. The principle advantage of this method are its adaptive capability, its low output noise and its low signal distorsion. The operating principle is shown in Figure 3.7.

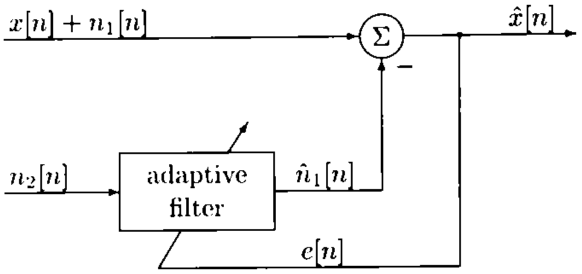


Figure 3.7: Principle of adaptive noise cancellation

The signal $x[n]$ corrupted by the additive noise $n_1[n]$, uncorrelated with the signal, is picked up by the primary microphone. The reference microphone receives a noise $n_2[n]$ correlated with the noise $n_1[n]$, but uncorrelated with the signal. The noise $n_2[n]$ is then adaptively filtered to produce an output $\hat{n}_1[n]$ which, subtracted from the primary input, will minimize the power of the output signal. In [Widrow et al. 75] it has been shown that minimizing the output power causes the output signal $\hat{x}[n] = x[n] + n_1[n] - \hat{n}_1[n]$ to be perfectly noise-free.
The adaptive filtering of the reference signal is necessary because of the possible delay between the arrival of the interference at the two microphones. The second reason is that the two microphones may pick up different versions of the noise, e.g. the noise at the primary input may be subject to echoes and/or spectrally variable attenuation [O'Shaughnessy 89].
The adaptive filter usually is a FIR filter of an order depending on the distance between the two microphones. Because of its simplicity, the LMS algorithm is the most widely used adaptive algorithm. The filter coefficients are updated to minimize the least-mean-square of the error signal $e[n]$. For positive SNRs, the adaptation constant μ of the LMS algorithm should be chosen [Van Compernolle 92] such that

$$\mu < \frac{0.1}{L}\max(P_{x+n_1}, P_{n_2}) \tag{3.36}$$

L is the number of taps of the FIR filter, P_{x+n_1} and P_{n2} are the power of the noisy primary signal and the power of the reference input, respectively. In this case the stability and convergence of the adaptation algorithm are guaranteed.
Adaptive noise cancellation does not result in significant SNR improvements in mobile systems because of the difficulty of satisfying simultaneously two fundamental assumptions [Liberti et al. 91]. The first requirement is that speech should be detected only by the primary microphone. Therefore the quality

of the noise reference is the most important issue in adaptive noise cancellation. Any speech detected by the reference microphone will be filtered and subtracted from the primary signal, thus reducing speech quality. Signal leakage into the noise reference leads to unacceptable distorsions in the filtered signal. In [Widrow & Stearns 85] it has been shown that for the case of speech captured by the reference microphone, the signal-to-noise density ratio[3] at the output of the adaptive noise canceller, $SNR_{out}[z]$, is given by

$$SNR_{out}[z] = \frac{1}{SNR_{ref}[z]} \tag{3.37}$$

Here, $SNR_{ref}[z]$ represents the ratio of the power spectra of the speech signal and the noise at the reference microphone. This means that the signal-to-noise density ratio at the output is the reciprocal, at all frequencies, of the reference input signal-to-noise density ratio. The process described by Eq. (3.37) is called *power inversion.*

Hence an almost signal-free noise reference is a must. The second requirement is that acoustic noise measured by the reference has to be very highly correlated to the noise from the primary input. The measure of noise cancellation $\psi(\omega)$ depends on the coherence function $\gamma(\omega)$ between the primary and reference signals [Goulding & Bird 90]:

$$\psi(\omega) = \frac{1}{1- \mid\gamma(\omega)\mid^2} \tag{3.38}$$

In [Dal Degan & Prati 88] a car environment is considered and the possibility of obtaining a suitable reference signal is investigated. Due to the spectral coherence of the noise in a car interior, a distance of less than 5 cm between the two microphones has to be chosen. This distance will permit a cancellation of at least 90% of the noise energy. However, with microphones placed at such a distance it is impossible to prevent the speech from entering both microphones.

On the other hand, if the two microphones are positioned at a distance greater than 50 cm, noise reduction will be performed only at very low frequencies (engine noise). In [Goubran et al. 90] it is found that the optimum location of the secondary microphone in a car depends very much on the driving conditions. Therefore the use of a parallel adaptive filter structure with more than one reference microphones is recommmended. It is possible to select different secondary microphones for every driving condition or to form the error signal as a combination of the error signals from the secondary microphones placed in different locations.

[3] The signal-to-noise density ratio is defined as the ratio of signal power density to noise power density and is a function of frequency [Widrow & Stearns 85].

3.2.2 Adaptive Beamforming

Speech beamforming can be applied when multiple noisy measurements are available. Considering the characteristics of speech and the special operation conditions, algorithms developed for beamforming in radar or sonar processing were tuned for speech applications [Van Compernolle 92].

The simplest beamformer is the *sum and delay beamformer* (Figure 3.8). The operating principle is the following: an array of M microphones provides a set of M noisy signals which are delayed and summed to produce a useful output [Ferrara & Widrow 81]. It is assumed that the signal components are correlated from channel to channel and that the noises are mutually uncorrelated and also uncorrelated with the signals. Considering that the angle θ of the speech signal incident on the array is known, the array may be *steered* or *beamformed* to the source by appropriately delaying the input signals from the other microphones. The delay time depends on the distance d_{mic} between the microphones and the angle θ [Silverman 87].

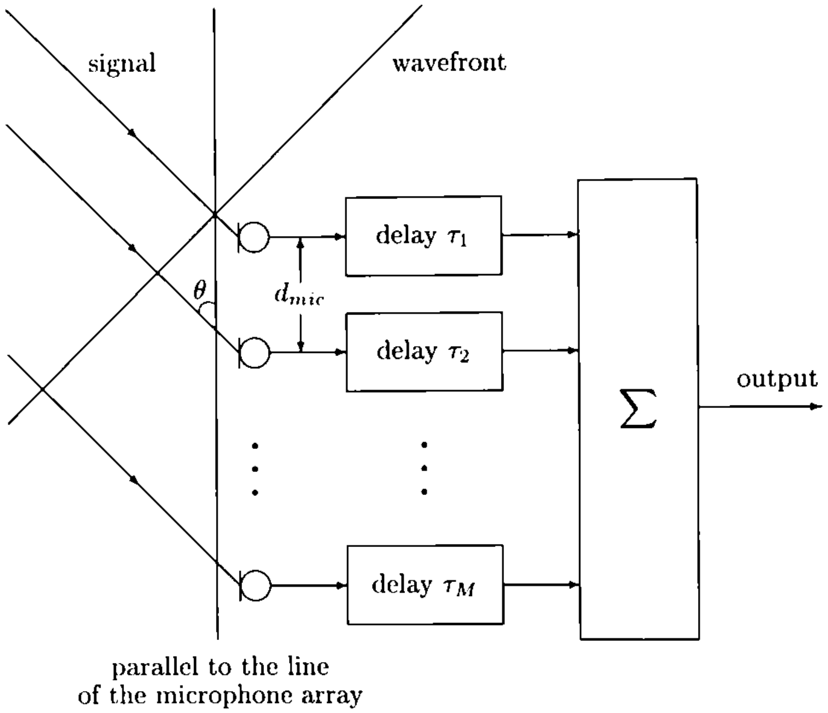


Figure 3.8: Delay and sum beamformer

Thus, adding the noisy signals yields an array output having a signal-to-noise ratio much improved over that of a single channel.

The delay and sum beamformer is very robust, errors in the estimate of delay times reduce the gain in SNR but cause only little distorsions. The great disadvantage of this beamformer is that the SNR-gain is limited and slowly increases with the number of microphones used. To obtain a gain of A dB over the single sensor method a number of $M = 10^{A/10}$ sensors is needed [Campbell 93]. In other words, a number of M microphones will give a theoretical maximum SNR-gain of $A = 10\, log_{10} M$ dB. In reality this gain will be smaller because of the background noise and the reverberation. Thus very large microphone arrays with huge hardware costs have to be designed.

An alternative to the delay and sum beamformer are the adaptive beamformers, which can be effective with a small number of microphones (in general 2 or 4) but require much higher computational investments.
The *constrained adaptive beamformer* takes advantage of potential correlation in the noise sources and thereby permits additional noise suppression by postprocessing in the form of an adaptive noise canceller. In the two-channel version of the *Griffiths-Jim beamformer*[4], an additional difference signal between the two inputs is computed, which is later used as noise reference in the noise canceller [Van Compernolle 92]. If phase alignment between both channels is perfect, the difference signal is an ideal noise reference. The assumption of a speech-free noise is not realistic because of the ever present reverberation and delay measurement error. In this case, speech will be present in the difference signal and the adaptive noise canceller will suppress speech as well. A speech detector which halts adaptation during periods with speech could minimize this problem.
For speech applications, the two channel adaptive beamformer has extensively been studied in the literature.
In [Faucon et al. 89] and [Faucon & Tazi Mezalek 90] two methods of noise reduction are presented which are based upon the assumption that speech signals as well as noises in the two observations are strongly correlated. Both structures consist of two stages:

- in the first stage a transfer function between the speech signals (first method) or the noises (second method) is identified. The learning of the transfer functions is done in absence of the noise or the signal, respectively. These transfer functions are then assumed to be stationary.

- the second stage performs the noise cancellation.

Due to the weak coherence between the noises, which depends on the distance between the microphones [Le Bouquin & Faucon 90] and also on their location and nature, the identification method of the noise transfer function was considered to be inefficient.

[4] The *Griffiths-Jim beamformer* is a special case of the more general *Frost beamformer* [Widrow & Stearns 85]

The (magnitude squared) coherence function, a frequency domain measure of correlation between two signals $y_1[n]$ and $y_2[n]$

$$\begin{aligned} y_1[n] &= x_1[n] + n_1[n] \\ y_2[n] &= x_2[n] + n_2[n] \end{aligned} \tag{3.39}$$

with n representing the discrete time index, is defined as follows:

$$C_{y_1y_2}[k] = \frac{|\gamma_{y_1y_2}[k]|^2}{\gamma_{y_1y_1}[k]\,\gamma_{y_2y_2}[k]} \tag{3.40}$$

where k represents the frequency bin, $\gamma_{y_1y_2}[k]$, $\gamma_{y_1y_1}[k]$ and $\gamma_{y_2y_2}[k]$ represent the *cross power densities* and the *power spectral densities* of the signals $y_1[n]$ and $y_2[n]$, respectively. The coherence function attains its maximum of 1 when the two signals are correlated, it is zero for uncorrelated signals. In practical situations the coherence function will vary between these two limits and determines, for each frequency, the percentage of signal energy coming from correlated sources.
The coherence function is observed to give important information in distinguishing useful signal from disturbing noise [Le Bouquin & Faucon 90], thus allowing a distinction between speech and noise. In a car environment, the coherence between the speech signal is almost 1, over all frequencies, while the coherence between noises decreases with the frequency and the distance between the two microphones [Dal Degan & Prati 88]. A method of speech enhancement which filters one observation y_1 using the coherence function elevated to a power δ is presented in [Le Bouquin & Faucon 92] and [Le Bouquin-Jeannès et al. 94]. The block diagram of the algorithm is presented in Figure 3.9.

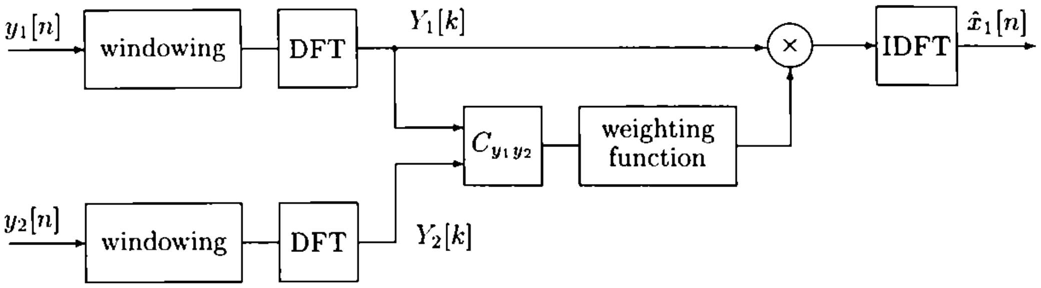


Figure 3.9: Speech enhancement using the coherence function

The coherence function has the task of turning off uncorrelated signals and passing through correlated signals. For this, two thresholds S_{min} and S_{max} are considered:

- if $C_{y_1y_2}[k] \geq S_{max}$, speech is considered to be predominant which must be passed

- if $C_{y_1y_2}[k] \leq S_{min}$, it is assumed that only disturbing noises are present, which must be eliminated
- if $S_{min} < C_{y_1y_2}[k] < S_{max}$, it is assumed that both signal and noise are present and therefore one observation e.g. $y_1[n]$ will be weighted by a function of $C_{y_1y_2}[k]$

The following equations describe the algorithm:

$$\hat{X}_1[k] = \begin{cases} C_{y_1y_2}[k]^{\delta}\, Y_1[k] & if \quad S_{min} < C_{y_1y_2}[k] < S_{max} \\ Y_1[k] & if \quad C_{y_1y_2}[k] \geq S_{max} \\ (S_{min})^{2\delta}\, Y_1[k] & if \quad C_{y_1y_2}[k] \leq S_{min} \end{cases} \tag{3.41}$$

The parameter δ allows a more selective filtering, an increased value of δ leads to a rigid filtering.
After finding $\hat{X}_1[k]$, the useful speech signal $\hat{x}_1[n]$ will be deduced.
The drawback of this system is the restrictive hypothesis that the noises must be decorrelated, otherwise the recognition system won't give good results. Sometimes some slight musical tones can be heard in the enhanced signal.

The problem of musical tones at low input signal-to-noise ratio also appears in the two microphones noise reduction algorithm from [Martin & Vary 94]. The main source of these musical tones is considered to be the residual correlation at low frequencies. These artefacts can be avoided if the noise reduction system processes only frequency components that lie above a minimum normalized frequency

$$\Omega_{min} = \frac{2\,\pi\, c}{d_{min}\, F_s} \tag{3.42}$$

where d_{mic} is the microphone distance, c the speed of sound and F_s stands for the sampling frequency. Knowing that

$$\Omega = \frac{2\,\pi\, f}{F_s} \tag{3.43}$$

means a frequency f_{min} of 850 Hz for a microphone distance of 40 cm. This constraint requires either a large microphone distance, which is not realizable, or the use of highpass filtering. The frequencies below the cutoff frequency of the highpass filter must be processed by some other noise reduction method or simply bypassed, without any processing.

Chapter 4

Digital Replica of the Loudspeaker-Room-Microphone-System

When designing acoustic echo cancellers and noise reduction systems for the mobile environment, certain specific considerations have to be taken into account. The vehicle interior, the operating conditions and the GSM requirements must be well known.

The *Loudspeaker-Room-Microphone-System (LRMS)* will be approximated by an adaptive filter because of the time-variant nature of the vehicle interior. The acoustic echo compensator must work adaptively, which means that it has to track any modifications of the LRMS by itself and as rapidly as possible. Its task is the estimation of the echo signal $\hat{y}[n] \simeq y[n]$ by generating the copy of the LRMS, filtering the loudspeaker signal with this replica and then subtracting the estimated echo from the microphone input signal.

Considering the small internal volume and the attenuation produced by the upholstery and the passengers, the length of the echo path in cars is between 30 ms [Goulding & Bird 90] and 60 ms [Armbruster et al. 91]. These values can be confirmed by the measurement of the impulse response in a middle size car (Opel Vectra limousine) performed during the development phase of the combined system presented in this thesis. Figure 4.1 shows the impulse response measured in a car[1]. With the sampling rate being 8,000 Hz, the 256 coefficients correspond to an impulse response length of 32 ms. The details of the filter specification involved in this procedure depend on two choices that have to be made:

- the filter type
- the type of statistical criterion used for the optimization

[1] The measurement procedure is described in Annex F.

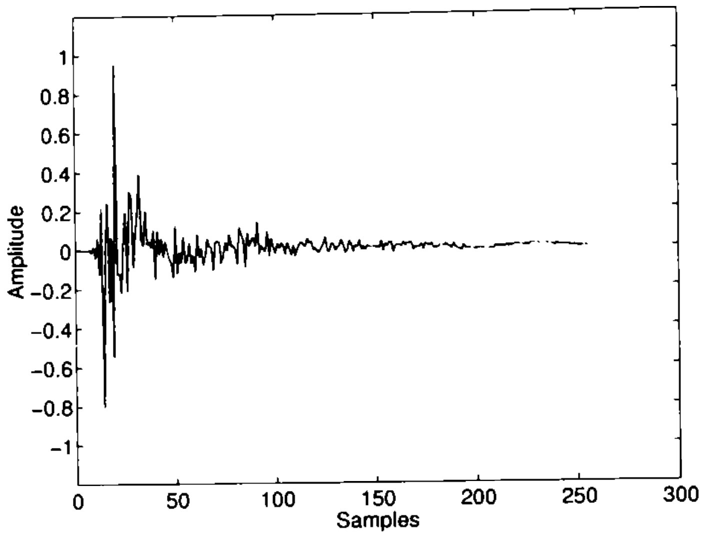


Figure 4.1: Impulse response measured in a car

4.1 Choice of the Filter Structure

The choice of a structure for the filtering process has a profound effect on the operation of the algorithm as a whole. The decision for a nonrecursive or a recursive filter design will be dictated by practical considerations.
The output of a *nonrecursive digital filter* depends on the current and one or more previous input samples. Such a filter implements the convolution sum directly, and the coefficients are simply equal to successive terms in its impulse response. Since the number of coefficients must be finite, a nonrecursive filter is also referred to as *FIR (finite impulse response)* filter.
As the transfer function of the FIR filter is specified in terms of z-plane zeros only, there is no danger that inaccuracies in the coefficients may lead to instability. Thus, the nonrecursive filter is inherently stable. Because of its finite impulse response, the FIR can be made symmetrical in form, which leads to an ideal linear phase characteristic, equivalent to a pure time delay of all frequency components passing through the filter. There is no phase distortion.

The output of a *recursive digital filter* depends on one or more previous output values, as well as on inputs, i.e. it involves both feedforward and feedback. In most cases a recursive filter has an *infinite impulse response (IIR)*. Although the impulse response decays towards zero, it theoretically continues forever. Assuming the filter is causal, this means that the impulse response cannot be symmetrical in form. Therefore the filter cannot display a pure linear phase characteristic. From the DSP point of view, the great advantage of the IIR filters is their computational economy. In average, a particular specified filter characteristic can be obtained with less than an 8-th of the number of filter coefficients [v.Zitzewitz 89] compared to its nonrecursive filter realization. However, there are two potential disadvantages of an IIR filter design:

- because of the feedback paths in the filter design, a recursive filter may become unstable with the result that it may oscillate
- recursive designs cannot generally provide the linear phase responses achieved by nonrecursive filters

The stability problem in IIR filters, e.g. in lattice filters, is manageable in both theoretical and practical terms, but when the filter is required to be adaptive additional problems have to be overcome. For this reason, in the majority of applications requiring the use of adaptivity FIR filters are preferred over IIR filters, even though the latter are less demanding in computational requirements. Owing to its versatility and ease of implementation, the FIR filter in its transversal structure juts out as an essential signal processing structure in a wide variety of applications.

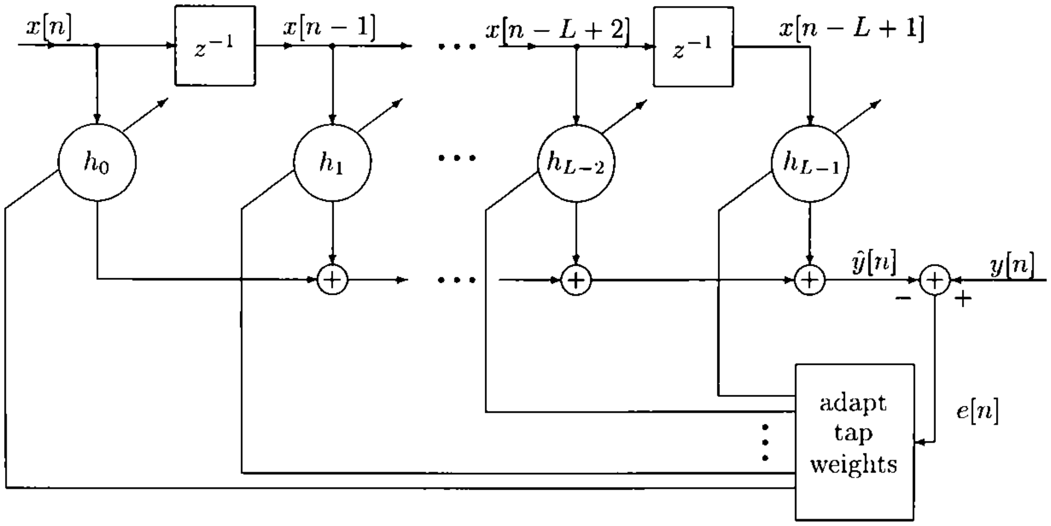


Figure 4.2: Adaptive transversal filter

The transversal filter, also referred to as a tapped delay line filter, consists of three basic elements, as depicted in Figure 4.2:

- unit-delay element, z^{-1}
- multiplier, h_k with $k = 0, 1, \ldots, L-1$
- adder

The number of delay elements L used in the filter determines the finite duration of its impulse response and is referred to as the filter order.

The delay elements are each identified by the unit delay operator z^{-1}. In particular, when z^{-1} operates on the input $x[n]$, the resulting output is $x[n-1]$. The role of each multiplier in the filter is to multiply the tap input to which it is connected by a filter coefficient called *tap weight*. Thus, a multiplier connected to the k-th tap input $x[n-k]$ produces $h_k\ x[n-k]$, where h_k is the respective tap weight and $k = 0, 1, \ldots, L-1$.
The role of the adders in the filter structure is to sum the individual multiplier outputs and produce an overall filter output.
For the transversal filter in Figure 4.2, the filter output is given by

$$\hat{y}[n] = \sum_{k=0}^{L-1} h_k\ x[n-k] \tag{4.1}$$

Equation (4.1) is called a finite convolution sum in the sense that it convolves the finite duration impulse of the filter, h_k, with the filter input $x[n]$ to produce the filter output $\hat{y}[n]$. The transversal filter is linear, i.e. the output is a linear function of the input samples and operates in discrete time, which enables the filter to be implemented using digital hardware/software.

The requirement for an adaptive transversal filter is to make the estimation error $e[n]$ defined as the difference between the desired response $y[n]$ and the filter output $\hat{y}[n]$ as small as possible, in some statistical sense. To achieve this, the filter coefficients will be made time-variant, $h_k[n]$ denoting the value of coefficient h_k at time instant n.

4.2 Choice of the Optimization Criterion

The choice of a statistical criterion for optimizing the filter design is influenced by mathematical tractability. Basically there are two distinct approaches to the development of linear adaptive filtering algorithms, depending on the cost function they minimize:

- the stochastic gradient approach, which minimizes the *mean-square error*, i.e. the mean-square value of the difference between the desired response and the actual filter output

- the least squares estimation, which minimizes the *sum of error squares*, where the error is defined as the difference between the desired response and the actual filter output

As a result of the discussions about the preferred filter structure from the previous section, the tapped delay line or transversal filter will be used as the structural basis for implementing the linear adaptive filter.

For the case of stationary inputs, the cost function for the stochastic gradient approach is defined as the mean squared error, i.e. the mean square value of the difference between the desired response and the transversal filter output:

$$J[n] = E[|\,e[n]\,|^2] \tag{4.2}$$

This cost function is a second order function of the tap weights in the transversal filter. The dependence of the mean squared error on the unknown tap weights may be viewed in the form of a *multidimensional paraboloid* with a uniquely defined bottom. This paraboloid is referred to as the *error performance surface.* The tap weights corresponding to the minimum point of the surface define the optimum Wiener solution.
The recursive algorithm for updating the tap weights of the adaptive transversal filter is known as the least mean square (LMS) algorithm, the essence of which may be described as follows:

$$\begin{pmatrix}\text{update value}\\ \text{of tap-weight}\\ \text{vector}\end{pmatrix} = \begin{pmatrix}\text{old value}\\ \text{of tap-weight}\\ \text{vector}\end{pmatrix} + \begin{pmatrix}\text{learning-}\\ \text{rate}\\ \text{parameter}\end{pmatrix} \begin{pmatrix}\text{tap-}\\ \text{input}\\ \text{vector}\end{pmatrix} \begin{pmatrix}\text{error-}\\ \text{signal}\end{pmatrix}$$

where the error signal is defined as the difference between some desired response and the actual response of the transversal filter produced by the tap input vector.

The most important member of the family of stochastic gradient algorithms is the LMS algorithm. The LMS algorithm is simple and yet capable of achieving satisfactory performance under the right conditions. Its major limitations are the slow rate of convergence and its sensitivity to variations in the condition number of the correlation matrix of the input signal, thus making it not suitable for speech applications.
In a nonstationary environment, the orientation of the error-performance surface varies continuously with time. In this case, the LMS algorithm has the additional task of continually tracking the bottom of the error-performance surface. This is possible, provided that the input data vary slowly compared to the learning rate of the LMS algorithm.

The second approach to the development of linear adaptive filtering algorithms is based on the method of least squares. According to this method, a cost function is minimized that is defined as the sum of error squares, where the error or residual is itself defined as the difference between some desired response and the actual filter output

$$J[n] = E[\sum_{i=i_1}^{i_2} |\,e[i]\,|^2] \tag{4.3}$$

i_1 and i_2 define the index limits at which the minimization occurs. The recursive algorithm for updating the tap weights of the adaptive transversal filter is known as the RLS algorithm, its derivation relying on the matrix inversion lemma. An important feature of the RLS is its much better convergence rate compared to that of the simple LMS algorithm. The limitations of the RLS algorithm include lack of numerical stability and an increased computational complexity. Aiming at the reduction of computational effort, the fast RLS algorithms were developed.

Within the context of this thesis, several of the fast RLS algorithms in the direct linear transversal form were implemented and their behaviour was studied. It was found that the numerical stability of these algorithms was a serious problem, most of them getting unstable after just a few minutes. Reinitialization with certain fixed values led to a reduction in convergence speed. The algorithms considered were the *Fast Kalman (FK)* algorithm, the *Stabilized Fast Kalman (SFK)*, the *Covariance Fast Kalman (CFK)*, the *Fast Transversal Filter (FTF)*, *Stabilized Fast Transversal Filter (SFTF)* and the *Corrected Fast Transversal Filter (CTFT)*. A general description of these algorithms can be found in [Schütze & Ren 92]. Deviations from the exact RLS algorithms and numerical instabilities due to unpredictable round-off errors are the main disadvantages of the fast RLS algorithms. Because of these and the additional stabilizing efforts that would have been necessary, it was decided not to use RLS algorithms in the acoustic echo cancellation.

Comparing the tracking behaviour of linear adaptive filters it can be stated that the stochastic gradient algorithms, such as the LMS algorithm, exhibit good tracking behaviour. This is due to the fact that they are model independent. In contrast, RLS algorithms are model dependent and therefore their tracking behaviour may be inferior to that of a member of the stochastic gradient family.

Relating to the realities of the acoustic echo compensation application, the choice has to be made in accordance with an optimum of computational cost, performance and robustness. These criteria inevitably lead to the decision of using a stochastic gradient algorithm.

4.3 Requirements for GSM Handsfree

Special conditions concerning the GSM system requirements are not to be neglected in the design procedure. The most important requirement in the GSM acoustic echo cancellation and noise reduction task is the processing time. The span of time including the *round trip delay*[2] and the processing time of the handsfree stages shall not exceed 143.9 ms + 39 ms, i.e. the acoustic cancellation and noise reduction functions have to be performed in no more than 39 ms.

[2] The round trip delay represents the sum of mobile station speech delay in uplink and downlink directions and shall not exceed 143.9 ms [GSM Rec. 03.50 96].

The GSM recommendations[3] specify performance characteristics and values, which acoustic echo cancellers must comply with, and methods to verify these performances. The most important characteristics are

- the *convergence time*, describing the convergence behaviour of the acoustic echo canceller which should be at least 1 s for an *Echo Return Loss Enhancement (ERLE)*[4] of at least 20 dB

- the *Terminal Coupling Loss (TCL)*, defined as the overall attenuation of the echo resulting from the the acoustic coupling of the terminal combined with the effect of the echo canceller [GTR SMG 97]. The TCL is the sum of the coupling loss between loudspeaker and microphone and the ERLE, which measures the intrinsic efficiency of the acoustic echo canceller. The TCL for the handsfree mobile station [GSM Rec. 03.50 96] shall be 40 dB at the nominal setting of the volume control in quiet background conditions and at least 33 dB at the maximum user selectable volume.

Automobiles both create and operate in a noisy environment. Handsfree telephones thus suffer from intrusive noise due to wind, fan or car engine. The noise field may appear diffuse or, due to the superposition of radiation from discrete sources, moving and statistically nonstationary. The speech/noise ratio in car environments can be as low as 0 dB, which means that the performance claim for noise reduction systems is extremely high.

In [GTR SMG 97] some values are given for the reduction of background noise in cars. Thus, an attenuation of 25 dB seems to be necessary to obtain good or excellent quality level in idle situation, with turned off engine. At a driving speed of 90 km/h, 15 dB of attenuation ensures a satisfactory quality level and a minimum of 12 dB at 130 km/h is desirable to assure an acceptable quality level.

4.4 Conclusions

The exact knowledge of the environmental conditions in which GSM car handsfree systems operate is crucial in the development of combined acoustic echo cancellation and noise reduction systems. As the acoustic echo canceller has the task of suppressing the echo generated by the loudspeaker-room-microphone system, first a digital replica of the LRMS must be defined. Because of the time-variant nature of the vehicle interior, the LRMS will be approximated by an adaptive filter.

[3] According to the *International Standardization Organization (ISO)* a *recommendation* is a binding document which contains legislative, regulatory or administrative rules and which is adopted and published by an authority legally vested with the necessary power [Walke 98].

[4] The definition of the ERLE is given in Chapter 5, section 5.3.1

Weighing the advantages and disadvantages of recursive and nonrecursive filters, the tranversal FIR filter in the linear direct form was proposed to be used because of its robustness and ease of implementation. IIR filters are much more difficult to handle, especially when the filter inherent feedback has to be combined with the adaptation process.

The second decision that had to be made in this Chapter was that concerning the adaptation algorithm. Because of the problem of numerical instability and increased computational load, the RLS approach will not be further pursued. Therefore it was suggested to perform a stochastic gradient algorithm which is numerically robust under well defined conditions for the convergence factor μ. It is also computationally less demanding than its competitor from the least-squares algorithms.

Based on this replica of the LRMS consisting of transversal FIR filter and stochastic gradient algorithm, an estimate of the room impulse response will be provided by the AEC system and the acoustic echo can be compensated by subtracting this adaptive estimate from the microphone signal.

Considering the GSM requirements on the handsfree system, the most important is the time constraint of 39 ms. This interval is specified by the GSM recommendations as being the worst case acceptable for processing delays associated with acoustic echo cancellation and noise reduction.

Chapter 5

Combined System of Acoustic Echo Cancellation and Noise Reduction

The goal of a combined system is to merge an acoustic echo canceller and a noise reduction system in a symbiosis in order to get a near-end speech signal with minimum distorsions and low levels of acoustic echo and background noise.

In the mono-channel approach the observation $y[n]$ received on the microphone is composed of

- a near-end speech signal $s[n]$ to be transmitted
- an echo $e[n]$ due to the signal $x[n]$ emitted by the loudspeaker
- and a background noise signal $n[n]$.

$s[n]$, $n[n]$ and $e[n]$ are additive and uncorrelated so that it can be written:

$$y[n] = s[n] + e[n] + n[n] \tag{5.1}$$

The signal $x[n]$ coming from the loudspeaker is correlated with the echo $e[n]$ and is used as reference input for the acoustic echo canceller (AEC).
The objective is to find an optimal structure in the sense of a minimal mean-square error by combining acoustic echo cancellation and noise reduction. Thus a good estimate $\hat{s}[n]$ of the near-end speech $s[n]$ is obtained. In the ideal case the near-end speech would be transmitted without any distortions and attenuations, while any acoustic echo and background noise would be suppressed.

A combined system of echo cancellation and noise reduction was first presented in [Yasukawa 92]. This system is a cascaded structure of two adaptive filters, the first one performing a noise reduction using a noise reference while

the second one operates as an acoustic echo canceller. The noise reduction filter is cascaded with the lower of the two bands of the AEC. The adaptation algorithm used for updating the filter coefficients of both filters is the normalized LMS.
As the system tends to become unstable if the filters are adapted simultaneously, a speech detector is used to control the adaptation. Thus, the noise reduction filter coefficients are updated only during the absence of near-end and far-end talk. The AEC coefficients are adapted when only the far-end speaker is active. During the AEC adaptation the noise reduction filter coefficients remain unchanged.

5.1 Combined System Structure

Considering speech as short-time stationary processes, the determination of the optimal filter in the sense of minimum mean-square error [Ayad & Faucon 95], [Le Bouquin-Jeannès et al. 96] leads to:

$$\hat{S}[k] = \frac{\gamma_{ss}[k]}{\gamma_{ss}[k] + \gamma_{nn}[k]} \left(Y[k] - \frac{\gamma_{xy}[k]}{\gamma_{xx}[k]} X[k] \right) \tag{5.2}$$

where $\hat{S}[k]$, $X[k]$ and $Y[k]$ represent the spectra of the signals $s[n]$, $x[n]$ and $y[n]$, respectively. $\gamma_{ss}[k]$, $\gamma_{nn}[k]$ and $\gamma_{xx}[k]$ are the power spectral densities of $s[n]$, $n[n]$ and $x[n]$. γ_{xy} represents the cross spectral density between the observations $x[n]$ and $y[n]$.
This equation shows [Ayad et al. 96], that there are two steps involved in the optimal structure presented above:

1. the echo is estimated by applying a filtering on the reference $x[n]$ with the transfer function given by
$$\gamma_{xy}[k] \,/\, \gamma_{xx}[k] \tag{5.3}$$
and then subtracting the filter output from the microphone observation. Thus, the term in brackets in Eq. (5.2) is the part of acoustic echo cancellation. For an ideal echo canceller, near-end speech and noise are transmitted with no change and the echo canceller output would be echo-free, i.e. $s[n] + n[n]$.

2. the noise is reduced by a Wiener filter with the following gain function:
$$\gamma_{ss}[k] \,/\, (\gamma_{ss}[k] + \gamma_{nn}[k]) \tag{5.4}$$

5.1.1 Acoustic Echo Cancellation Preceding Noise Reduction

As just stated, the optimal structure is composed of two cascaded optimal filters, where the first one performs the acoustic echo cancellation and the second one is a noise reduction system. The output of the AEC, ideally $s[n]+n[n]$, is subjected to Wiener filtering. This structure will be called *AEC+NR* and is represented in Figure 5.1.

In this structure the identification operation of the AEC system is disturbed by the omnipresent background noise and the near-end speech in double-talk situation. In [Guelou et al. 96] it is shown that the performance of the AEC+NR structure depends very much on the intrinsic behaviour of the implemented adaptation algorithm. An adaptation algorithm with improved robustness to noise and double-talk will perform much better than the generally used NLMS algorithm, known for its lack of robustness to noise.

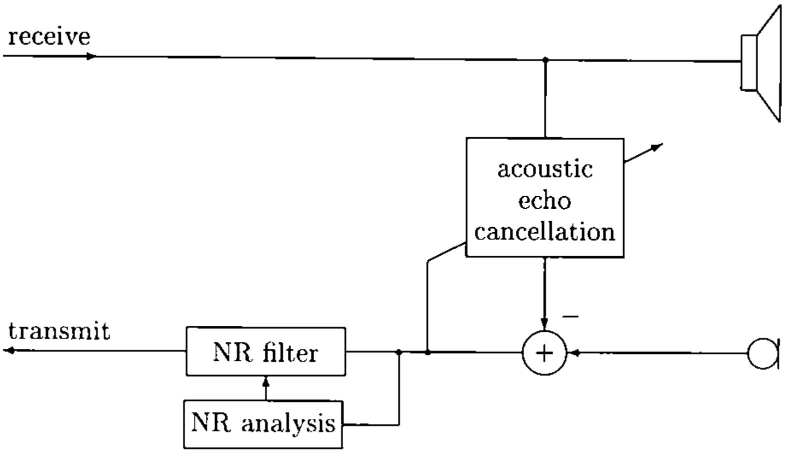


Figure 5.1: AEC + NR structure

As the near-end signal may be distorted at the output of the AEC, it was proposed [Ayad & Faucon 95], [Le Bouquin-Jeannès et al. 96] to take the input signal for the noise reduction system from the microphone input $y[n]$ and not from the output of the acoustic echo canceller. In this new structure, presented in Figure 5.2, the NR filter and the AEC system are estimated simultaneously.

It was found that the distorsion brought by the NR system to near-end speech signal during double-talk periods is lower in the modified structure. However, in single-talk mode the standard AEC+NR structure yields a greater reduction in noise and echo. Thus a double-talk detector could help to optimize the performance in both modes.

5.1.2 Noise Reduction Preceding Acoustic Echo Cancellation

To reduce the noise influence on the AEC system, the NR system can be placed in front of the AEC, the elimination of the undesired background noise being

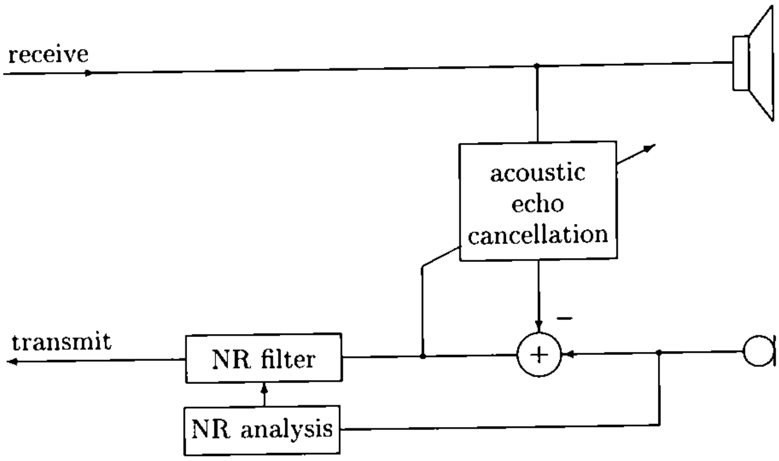


Figure 5.2: Modified AEC + NR structure

performed before the adaptation process of the AEC. This structure will be called *NR + AEC* and is presented in Figure 5.3.

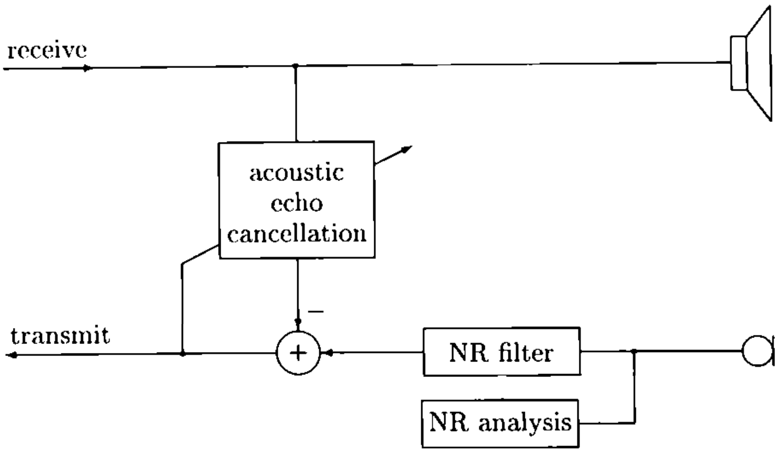


Figure 5.3: NR + AEC structure

The echo estimate in the NR+AEC structure is closer to the original echo [Ayad et al. 96], but the noise reduction operation also distorts the echo signal. The noise reduction system introduces a time-varying filter into the acoustic echo path. These distorsions can disturb the identification process [Guelou et al. 96]. The performance of this basic structure may be enhanced by introducing a copy of the noise reduction filter in the identification branch (Figure 5.4). This will reduce the non-linear distorsions [Martin & Vary 94]. However, because of the time-variant behaviour of the filter, the adaptation process of the AEC has to be executed each time the coefficients of the noise reduction filter change. This recomputation is increasing the complexity.

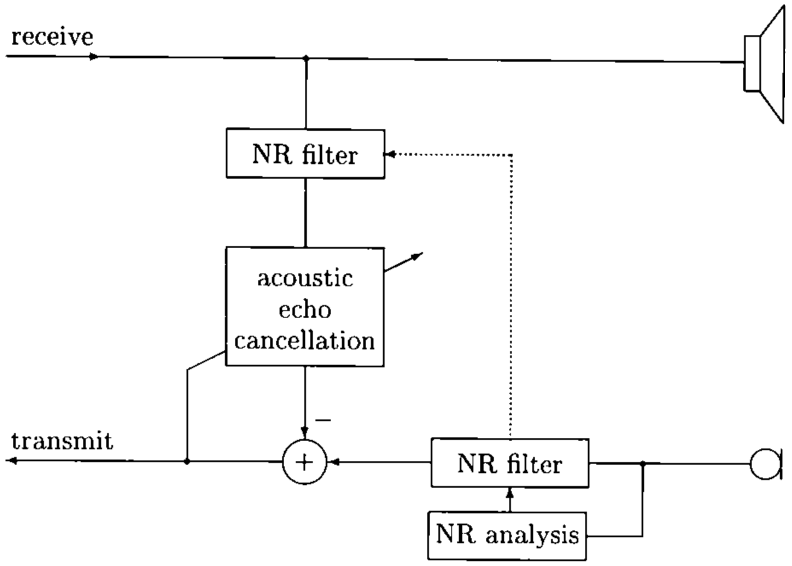


Figure 5.4: Modified NR + AEC structure

The limitations of combined systems are summarized in the so-called *speech enhancement trilemma* [Martin et al. 96]. The trilemma requires a compromise between the echo and noise reduction, the introduced distorsion of the near-end speech signal and the complexity of the overall system. Adding more microphones or more compensator coefficients will improve the echo compensation and noise reduction at the price of an increased complexity. On the other hand, keeping the complexity constant, a trade-off must be found between echo and noise reduction and the near-end speech distorsions.

5.2 The Proposed Combined System

After the analysis, implementation and test of the hitherto presented algorithms, a new system is proposed which works exclusively in the time domain. Because of the restrictions considering the implementational cost in a GSM mobile terminal, the intention was to use simple algorithms with reduced complexity. The noise reduction is considered in the approach using one microphone, the reference noise estimate being delivered by a voice activity detection algorithm. The great advantage of an entirely time domain implementation is the absence of delay, except for the processing time.
The combined system of acoustic echo cancellation and noise reduction consists of the following elements:

- an acoustic echo canceller based on the *Affine Projection Algorithm (APA)* and a transversal filter
- a far-end voice activity detector

- a double-talk detector
- a near-end voice activity detector
- a noise reduction system also based on the APA and a transversal filter

The block diagram of the proposed AEC-NR system is presented in Figure 5.5. A/D and D/A conversion blocks are considered to be included in the representation of the microphone and loudspeaker. The positioning of AEC+NR within the GSM communication chain is shown in Annex B, Figure B.2.
The main elements of the handsfree system are the acoustic echo canceller and the noise reduction system, which have to perform the task of eliminating the far-end speaker's echo and reducing the environmental noise inside the car cabin. The proposed algorithms will be presented in detail in sections 6.1 and 8.1. The system is completed by the use of speech detectors on the far-end and near-end signal which will be discussed in Chapter 7.

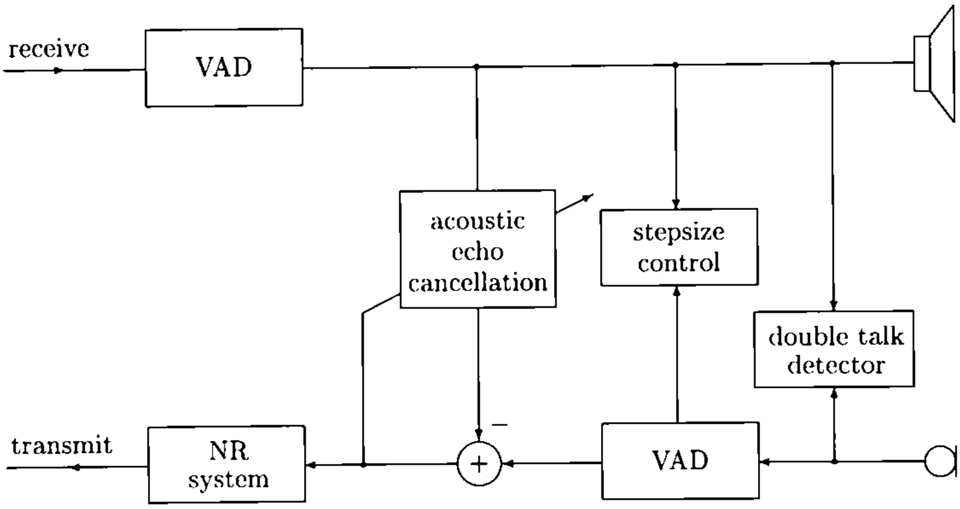


Figure 5.5: The proposed combined system

The role of the far-end VAD is that of detecting the presence of speech on the receive path. If speech is present, it will be emitted by the loudspeaker into the car interior, several times reflected and then picked up by the microphone. The disturbing acoustic echo can be eliminated by the adaptation and filtering processes of the acoustic echo canceller. In the absence of speech on the receive path, no adaptation and filtering is necessary. Therefore, when no far-end speech activity is detected, the operation of the AEC will be stopped.
In the presence of a reference signal on the far-end side, the decision on adaptation is handed over to the double-talk detector (DTD). If the DTD supplies the information that there is no double-talk, the AEC performs normally, adapting the coefficients of the transversal filter. If speech is considered to be present

at the near-end, the coefficients' adaptation process is stopped, yet filtering still continues. This is equivalent to setting the adaptation coefficient μ to zero, while the AEC continues to perform. Thus the last estimated echo will be subtracted from the microphone input signal, but the adaptation will proceed no longer.

For a more noise robust operation of the AEC algorithm, a near-end noise dependent stepsize control is additionally implemented. This control presented in section 6.4, varies the value of μ depending on the ratio between the loudspeaker signal power and the estimated noise power. In high background noise the adaptation coefficient will be close to zero, thus slowing down the adaptation process.
The near-end VAD is used to estimate the noise background, which is necessary for the stepsize control algorithm and the following noise reduction system. This noise information can be obtained during pauses in the speech flow, and will be continuously updated during these speech-free periods.

The combined system will be implemented on a DSP. Considering the block diagram of a GSM handy, the AEC+NR system will be running on the DSP unit of the GSM chipset, as presented in Annex C, Figure C.1. It will be positioned between the transmit path (before of the speech encoder) and the receive path (after the speech decoder) as shown in Figure B.2.
Another possibility of implementation consists in placing the AEC+NR function entirely into an independent supplimentary DSP housed in an extra handsfree module. In this case the echo- and noise-free speech samples will be connected to the input of the GSM handy and no further processing will be necessary in the GSM chipset. The advantage of this placement is the possiblity of implementing more powerful algorithms, but at the expense of additional hardware costs caused by the supplimentary DSP.
In the approach of this thesis, the first case will be considered with its subsequent consequences of less computational power and memory. The problem is that cost is a very important factor for mobile telephone manufacturers.

5.3 Objective Performance Evaluation

During the development of acoustic echo cancellers and noise reduction systems, it is very important to assess the system's performance under different operating conditions or with different configurations. These evaluations can be performed by subjective listening tests and objective quality measures. As the subjective tests are extremely time and resource consuming and not always reproducible, the objective measures are preferred during the development phase of a speech processing algorithm.

The ideal objective quality measure would need to assess all the levels of human speech processing [Quackenbush et al. 88], i.e. psychoacoustics, acoustic-phonetics, morphology, prosodics, syntax, semantics, linguistics and pragmatics. Such an objective measure which considers all the above mentioned items cannot be applied in practice. Generally, only comparisons between the original and the distorted signals will be done, over short intervals of 10 to 30 ms duration where the speech characteristics do not change considerably.
The objective measure must be easy to perform on the enhanced signal, it must be highly correlated to the results obtained by listening tests and it must be selected to suit the specific application [Gustafsson et al. 96]. In some applications the maximum possible intelligibility is desired, while for other applications the minimization of listener fatigue is the main objective, i.e. the enhancement of the naturalness and pleasantness of speech.

The evaluation of the acoustic echo cancellation and noise reduction can be performed with real conversational speech or with test signals, defined in [ITU-T P.501 96] and presented in Annex D. Since the test signals are standardized, the reproducibility of results is guaranteed.

5.3.1 Acoustic Echo Cancellation

The performance of an echo canceller can be given in terms of its dynamic and steady-state properties. The dynamic performance is described by the the rate of initial convergence, while the steady-state performance is given by the misadjustment of the adaptive filter after convergence.

Considering an acoustic echo compensator as shown in Figure 5.6 several objective measures for evaluating the effectiveness of the compensation algorithms can be defined. Here, L represents the filter length, $\hat{\mathbf{h}}[n]$ a set of time- varying filter coefficients and $\mathbf{h}$ describes a time discrete weighting function of the loudspeaker-room-microphone-system. n is used to indicate the discrete time instant $(n = 0, 1, 2, \ldots)$.

The steady-state performance can be objectively assessed by using the following measures:

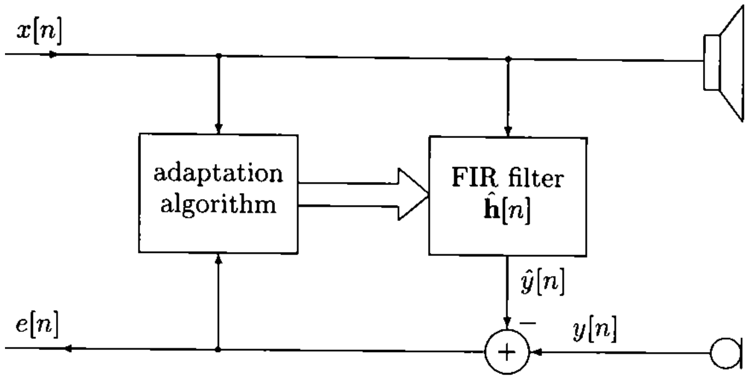


Figure 5.6: Time discrete structure of an acoustic echo canceller

- **relative error parameter norm** ($P[n]$) [Schütze 89]:

$$\begin{aligned} P[n] &= 20 \log \frac{\| \mathbf{h} - \hat{\mathbf{h}}[n] \|}{\| \mathbf{h} \|} \\ &= 10 \log \frac{(\mathbf{h} - \hat{\mathbf{h}}[n])^T (\mathbf{h} - \hat{\mathbf{h}}[n])}{\mathbf{h}^T \mathbf{h}} \\ &= 10 \log \frac{\sum_{k=0}^{L-1} (h_k - \hat{h}_k[n])^2}{\sum_{k=0}^{L-1} h_k^2} \end{aligned} \tag{5.5}$$

where n represents the time instant. This distance measure describes the degree to which the echo canceller of order L corresponds to the impulse response $\mathbf{h}$ of the loudspeaker-room-microphone system [Heitkämper & Walker 93]. It presents the following attributes:

  - the initial value is always 0 dB because at the beginning $\hat{\mathbf{h}}[0]$ is $\mathbf{0}$
  - $P[n] < 0$ means less residual echo than in the uncompensated case
  - $P[n] > 0$ means a deterioration, i.e. more residual echo is present than in the uncompensated case
  - for $n \to \infty$, $P[n]$ should ideally tend to $-\infty$ dB

- **Echo Return Loss Enhancement** (*ERLE*)
The ERLE gives a measure of the misadjustment of the adaptive filter. It is defined by the logarithmic expression of the ratio between the power of the desired signal $y[n]$ and the power of the difference signal $e[n]$ after compensation. The better the acoustic echo cancellation the larger the ERLE. Depending on the interval for which the power calculation takes place, the following two types of ERLE can be defined:

- **sliding window ERLE** ($ERLE_{slw_k}$)

$$ERLE_{slw_k} = 10 \log \frac{\sum\limits_{n=k-\lambda}^{k} y[n]^2}{\sum\limits_{n=k-\lambda}^{k} e[n]^2} \tag{5.6}$$

  where λ represents the constant length of a sliding window $t_\lambda = \lambda \cdot T$, with T standing for the sampling period. Considering a sampling frequency of 8,000 Hz, for white noise λ will be set to 800, which corresponds to a window length of 100 ms. When dealing with non-stationary speech signals, a window length of 1 second will be chosen, because of the pronounced ERLE variations during short speech pauses which cannot be confirmed by subjective listening. Taking a larger window lets the ERLE converge to the subjective impression of the echo attenuation.

- **segmental ERLE** ($ERLE_{seg_k}$)

$$ERLE_{seg_k} = 10 \log \frac{\sum\limits_{n=kM}^{kM+M-1} y[n]^2}{\sum\limits_{n=kM}^{kM+M-1} e[n]^2} \tag{5.7}$$

  where k ($k = 0, 1, 2, \ldots$) represents a segment of length M. The power calculation required by the ERLE is performed over short segments of speech, usually of 20 ms duration.

The described steady-state parameters can be evaluated both during single talk and double-talk periods.

The dynamic performance of the acoustic echo canceller can be objectively assessed by using the following measures [Naylor et al. 94]:

- **initial convergence time** (T_{ic})
  defined as the time needed by the acoustic echo compensator to attain the mean average value $ERLE_{mean}$ of the segmental ERLE calculated over the whole signal

- **the time to attain 10 dB of segmental ERLE** (T_{ic10dB})

When stationary signals are used, the adaptive filter will initially converge until it attains its mean misadjustment and then hold this value constant. As the statistics of speech are nonstationary, the adaptive echo canceller will be continuously tracking the varying echo path, thus no such definite convergence to the steady-state can be observed.

5.3.2 Noise Reduction

Speech enhancement systems usually consist of an adaptive filter in the signal path. As the phase of noisy signals is difficult to estimate, these filters are designed to modify only the amplitude spectrum and not the phase of the disturbed signal. Therefore, complete noise reduction is not possible and most noise reduction systems are distorting the speech signal more or less [Gustafsson et al. 96]. To estimate the improvement achieved by a speech enhancement system, the objective measures are calculated at the input and at the output of the noise reduction block. $x[n]$ represents the clean speech signal at time instant n, while $x_d[n]$ is considered to be the noisy signal at the input or the enhanced noisy signal at the output of the speech enhancement system.

The most commonly used objective measures for assessing the performance of a noise reduction system are:

- **segmental Signal-to-Noise Ratio improvement** (SNR_{seg})

  The *classical* SNR gives some indication of the quality of stationary systems [Quackenbush et al. 88] and is measured as

$$SNR = 10\,\log_{10}\frac{\sum_n x^2[n]}{\sum_n (x[n]-x_d[n])^2} \tag{5.8}$$

  As speech signals are nonstationary, the classical SNR is obviously not adequate for estimating speech quality. If the measurement described in Eq. (5.8) is taken over short segments of speech and then summed over all segments in that waveform, the result is a very good estimation of speech quality and is called *segmental* SNR. The segments where stationarity can be assumed are typically chosen to be 15 to 20 ms. The use of this segmentation permits an equal weighting of both loud and soft portions of the utterance. The same noise level may have different effect on the output signal, depending on the instantaneous input signal level.

  The segmental SNR is expressed as

$$SNR_{seg} = 1/N \sum_{i=0}^{N-1} 10\log_{10} \sum_{n=iM}^{iM+M-1} \left(\frac{x^2[n]}{(x_d[n]-x[n])^2}\right) \tag{5.9}$$

  with M representing the segment length and N the number of segments in the speech signal.

  The segmental SNR must be combined with a speech detector. Only the SNRs of the segments containing speech will be included in the sum in Eq. (5.9). Otherwise, any speech pause will give rise to a large negative signal-to-noise ratio which could appreciably bias the overall measure of the segmental SNR.

  It is important that the system is of linear phase, otherwise the segmental

SNR will not correspond to the perceived results. Any phase distorsion may reduce the $\mathrm{SNR_{seg}}$ significantly [Gustafsson et al. 96].
The gain G [Faucon & Le Bouquin-Jeannès 95] of a noise reduction system is obtained by subtracting the input segmental SNR from the output segmental SNR:

$$G = SNR_{seg_{output}} - SNR_{seg_{input}} \tag{5.10}$$

- **LPC spectrum matching measures**
  Speech enhancement algorithms can be also evaluated in the spectral domain. These measures are very sensitive to any changes in the spectral shape of the analyzed speech segment. The LPC based measures between the clean and the noisy/enhanced speech have been found to be very effective [Ahmed 89]. The better the spectrum matching between the clean and the processed speech signal, the better is the enhancement algorithm and the smaller the value of the spectral distance measure.
  For this type of measures, the clean and the noisy/enhanced speech waveforms are usually divided into analysis frames of 15 to 30 ms duration and a linear prediction analysis is done for each frame. The distance measure is computed from the results of the analysis.

Cepstral Distance
A cepstrum[1] computed from the predictor coefficients provides an estimate of the smoothed speech spectrum.
According to [Quackenbush et al. 88], it can be written:

$$\log\left(\frac{1}{A[z]}\right) = \sum_{k=1}^{\infty} c[k]\, z^{-k} \tag{5.11}$$

where $A[z]$ is the LPC model and $c[k]$ are the cepstral coefficients which can be computed from the predictor coefficients recursively:

$$n\, c[n] - n\, a[n] = \sum_{k=1}^{n-1} (n-k)\, c[n-k]\, a[k] \quad \text{for} \quad n = 1, 2, 3, \ldots \tag{5.12}$$

with $a[0] = 1$ and $a[k] = 0$ for $k > p$. n represents the time index. In Eq. (5.12) $a[k]$ are the predictor coefficients and p is the order of the LPC model.

[1] The *cepstrum* or *spectral function* of a speech signal is defined as the Fourier Transform applied to the logarithm of the Fourier Transform of the speech signal [Rowden 91]. The cepstrum is made up of a set of discrete cepstral coefficients, which are the output set of the final Fourier Transform process.

The cepstral distance, based on the cepstral coefficients is presented in [Le Bouquin et al. 93]:

$$d_{cep} = \sum_{n=1}^{2p} (c[n] - c_d[n])^2 \tag{5.13}$$

with $c[n]$ and $c_d[n]$ representing the cepstral coefficients corresponding to the speech signals $x[n]$ and $x_d[n]$ respectively.

Log Likelihood Ratio

The log likelihood ratio or *Itakura distance* is based on the dissimilarity between all-pole models of the clean speech signal and the distorted speech. It is assumed that over intervals of 15 to 30 ms speech can be represented by a p-th order all-pole model.

The log likelihood ratio compares two windowed speech signals, $x[n]$ and $x_d[n]$ and can be defined as [Quackenbush et al. 88]:

$$d_{LLR} = \log_{10}\left(\frac{\mathbf{a}_d \mathbf{R} \mathbf{a}_d^T}{\mathbf{a} \mathbf{R} \mathbf{a}^T}\right) \tag{5.14}$$

where $\mathbf{a}$ is the LPC coefficient vector $(1, -a[1], -a[2], \ldots, -a[p])$ for the original speech $x[n]$ and $\mathbf{a}_d$ the LPC coefficient vector for the distorted speech $x_d[n]$. $\mathbf{R}$ is the autocorrelation matrix for $x[n]$ with its elements defined as:

$$r[k] = \sum_{n=1}^{N-k} x[n]\, x[n-k] \qquad \text{for} \quad k = 0, 1, 2, \ldots, p \tag{5.15}$$

where N is the length of the frame used in the LPC analysis.

Itakura-Saito Distorsion Measure

Considering two spectral models $\sigma/A(z)$ and $\sigma_d/A_d(z)$ corresponding to the clean and distorted speech signals, respectively, the *Itakura-Saito* measure can be defined as follows [Gray & Markel 76]:

$$d_{IS} = \int_{-\pi}^{+\pi} \left[e^{V(\theta)} - V(\theta) - 1\right] \frac{d\theta}{2\pi} \tag{5.16}$$

with

$$V(\theta) = \log\left(\frac{\sigma^2}{|A(e^{j\theta})|^2}\right) - \log\left(\frac{\sigma_d^2}{|A_d(e^{j\theta})|^2}\right) \tag{5.17}$$

$A(z)$ is the inverse of the all-pole filter that models the spectral envelope of the sequence $x[n]$ and is defined as

$$A(z) = 1 - \sum_{i=1}^{p} a[i] z^{-i} \tag{5.18}$$

The Itakura-Saito measure appears in a number of different formulations, which are mathematically related under certain assumptions.
In [Gray & Markel 76] and [Le Bouquin et al. 93] the Itakura-Saito distorsion measure is defined as follows :

$$d_{IS} = \frac{\sigma^2}{\sigma_d^2}\,\frac{\delta}{\alpha} + \ln\frac{\sigma_d^2}{\sigma^2} - 1 \tag{5.19}$$

where

$$\delta = \mathbf{a}_d^T\,\mathbf{R}\,\mathbf{a}_d, \qquad \alpha = \mathbf{a}^T\,\mathbf{R}\,\mathbf{a} \tag{5.20}$$

$\mathbf{a}$ and $\mathbf{a}_d$ being the LPC coefficient vector for the original speech and the distorted speech, respectively.
$\mathbf{R}$ is the $((p+1)\times(p+1))$ input sample autocorrelation symmetric Toeplitz matrix, whose first row consists of $(p+1)$ autocorrelation values $r[k]$ with $k = 0, 1, 2, \ldots, p$. The gain terms of the models are denoted by σ and σ_d, respectively.

Taking into account that the perceptual frequency resolution is decreasing with increasing frequency, a frequency-weighted Itakura-Saito measure [Chu & Messerschmitt 82] can be defined as

$$d_{IS} = \int_{-\pi}^{+\pi} \left(e^{V(\theta)} - V(\theta) - 1\right) |\,W(e^{j\theta})\,|^2\,\frac{d\theta}{2\pi} \tag{5.21}$$

where the non-negative factor $|\,W(e^{j\theta})\,|^2$ weights low frequencies more heavily than high frequencies. Thus, the frequency weighted spectral estimation improves the accuracy of psychoacoustic representation of speech.

- **noise reduction factor** (R)
  Considering the case when noise and speech are known separately and the noisy signal is obtained by addition of the two signals, the noise reduction processing can be applied only to the noise. Thus the noise reduction performance of the algorithm [Le Bouquin et al. 93] can be measured. The filtered noise signal will be denoted by $n_f[n]$.
  The segmental noise reduction factor R will be computed as follows:

$$R = \frac{1}{N}\sum_{i=0}^{N-1} 10\,\log\frac{\sum\limits_{n=iM}^{iM+M-1} n[n]^2}{\sum\limits_{n=iM}^{iM+M-1} n_f[n]^2} \tag{5.22}$$

  where N is the number of segments of length M.

- **distorsion measure** (D)
  Making the same assumption as in the noise reduction factor computation,

the noise reduction procedure will be applied only to the clean signal $x[n]$. The distorted part $\varepsilon_f[n]$ may be written as

$$\varepsilon_f[n] = x_f[n] - x[n]. \tag{5.23}$$

Similarly to the noise reduction factor, the segmental distorsion D can be computed as follows [Faucon & Le Bouquin-Jeannès 95]:

$$D = \frac{1}{N} \sum_{i=0}^{N-1} 10 \log \frac{\sum\limits_{n=iM}^{iM+M-1} \varepsilon_f[n]^2}{\sum\limits_{n=iM}^{iM+M-1} x[n]^2} \tag{5.24}$$

5.4 Signals Used in the Experimental Part

During the experiments different speech and noise signals were used. The noise-free files included the following texts spoken by a male and a female speaker, respectively.

The male speaker's text is the following: "In the course of a December tour in Yorkshire, I rode for a long distance in one of the public coaches on the day before Christmas." It is stored in a data file having a length of 65,536 16-bit words, sampled with 8 kHz, i.e. the duration of the file is approximately 8.2 seconds.

The woman speaks the following text: "The simplest method is to mix the medicine with butter or some other grease and smear it on the nose of the animal from time to time; naturally, it will lick the grease off and in this way will swallow the medicine." This is a sentence of 12.3 seconds duration which, when sampled at 8 kHz, corresponds to a stored data file of 98,304 16-bit words.

The noisy files were constructed by adding noise to the clean speech signals. The car noises used in these experiments were recorded in driving cars:

- an Opel Astra 7 Caravan, with a driving speed of about 90 - 120 km/h on a normal road, with closed windows and the fan turned off. The record is approximately 11.4 seconds long. Its length is 91,058 16-bit words at a sampling frequency of 8 kHz.

- a BMW 540, moving at 90 - 120 km/h on a normal road, the windows were closed, the weather was rainy and the fan was turned off. The data file has a length of 134,0158 16-bit words. Sampling at 8 kHz, this means a duration of the data file of approximately 16.8 seconds.

These noise signals were available from the *CSDC2 Speech Database* created by the *Institute of Phonetics and Speech Communications of the University of Munich.*

White noise and the standardized the test signals were available from the CD-ROM being part of [ITU-T P.501 96].

When generating the noisy files, the duration of all files was truncated to that of the shortest sample file, i.e. to 8.2 seconds. The weighted noise file was added so as to obtain noisy files of different SNRs (-10 dB, 0 dB and +10 dB).
For generating the echo signal, necessary for measuring the acoustic echo cancellation performance, the speech signals from the data files were convolved with the impulse response measured in a car (see Figure 4.1) and attenuated by 20 dB.

Additionally, there were used true loudspeaker and microphone signals recorded in a car. The loudspeaker was located at the right of the navigator's feet, while the microphone's position was at the driver's left, on the A-pillar of the car. The speech signals were fed into the loudspeaker, and the echo signal was picked up by the microphone. Different driving conditions were simulated by feeding simultaneously the four car loudspeakers, placed in the doors, with earlier recorded car noise.
The double talk situation was also simulated. Therefore, an additional speech signal was emitted from a position on the driver's seat which would correspond to a near-end speaker's mouth location.

5.5 Conclusions

Acoustic echo compensation and noise reduction are two distinct functions applied in the field of handsfree operation. These separately working systems can be merged into a single combined system. The optimal structure in the sense of minimum mean-square error is given by the acoustic echo canceller preceding the noise reduction system.
Knowing this, a new combined system is proposed having its elements working entirely in the time domain. This has been considered because of the time constraint of 39 ms of processing time for both acoustic echo compensation and noise reduction imposed by the GSM specifications. A noise reduction system in the frequency domain, as it is usually considered, would need a Fourier Transform implementation. This would require already 32 ms of processing time, when the transform is performed over 256 samples. A noise reduction algorithm in the time domain operating on a sample-by-sample basis would be computationally less demanding and would save processing time.

For a more noise robust operation of the handsfree system, a near-end noise dependent stepsize control, a far-end and a near-end VAD, and a DTD will complete the combined system. The detailed description of the main elements of this system will be presented and discussed in the following chapters.

This Chapter ends with the presentation of some objective assessment procedures of AEC and NR performance and of the signals used in the experimental part of this thesis. As subjective listening test are very time and resource consuming and generally not reproducible, the use of objective measures is very important in the development phase of a combined system. The use of standardized test signals, as those presented in Annex D, has the advantage of supplying reproducible results.
The most important objective measures as well as the presented test signals will be used in the assessment of the new algorithms developed and presented in Chapters 6, 7 and 8.

Chapter 6

Adaptation in the Combined System

On the basis of the decision made in section 4.2 of Chapter 4, considering the optimization criterion for the filter design, a stochastic gradient algorithm will be used in the proposed combined system of acoustic echo cancellation and noise reduction. Thus, the cost function which will be minimized is the mean square value of the estimation error.

The block diagram of an adaptive filter is shown in Figure 6.1. The output of the unknown system to the input signal $x[n]$ will be $y[n]$, also called the desired signal.

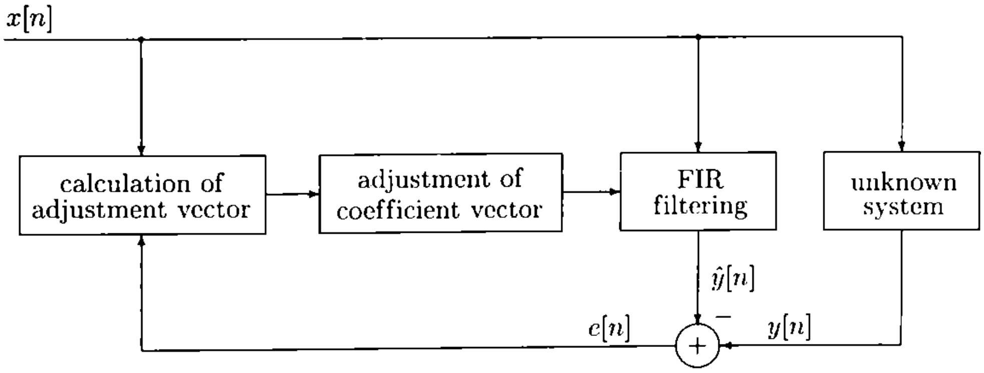

$x[n]$
calculation of adjustment vector
adjustment of coefficient vector
FIR filtering
unknown system
$\hat{y}[n]$
$e[n]$
$-$
$y[n]$
$+$

Figure 6.1: Block diagram of an adaptive filter

Assuming that the unknown system can be modelled by an FIR filter of length L with coefficients

$$\mathbf{h} = [h_1, h_2, \ldots, h_L]^T \tag{6.1}$$

the adaptation system has to estimate the unknown system by an FIR filter with coefficients

$$\hat{\mathbf{h}}[n] = [\hat{h}_1[n], \hat{h}_2[n], \ldots, \hat{h}_L[n]]^T. \tag{6.2}$$

n denotes the discrete time index.

For a sample-by-sample adaptation, the estimation filter or coefficient vector $\hat{\mathbf{h}}[n]$ is adjusted at every sample instance so as to make the estimation output $\hat{y}[n]$ close to the unknown system's output $y[n]$, i.e. a minimum for the estimation error $e[n]$. This is done by adding an adjustment vector $\Delta\hat{\mathbf{h}}[n]$

$$\hat{\mathbf{h}}[n+1] = \hat{\mathbf{h}}[n] + \mu[n]\,\Delta\hat{\mathbf{h}}[n]. \tag{6.3}$$

$\mu[n]$ is a time-variant scaling factor called stepsize, which controls the convergence speed and the amount of residual error.

From the variety of recursive algorithms, developed in the literature for the operation of linear adaptive filters, the selection of the preferred algorithm was determined by considering the rate of convergence, misadjustment, tracking behaviour, robustness and computational requirements. The decision fell upon the affine projection algorithm, an algorithm rediscovered a few years ago and implemented in its fast version, which makes it computationally less demanding compared to its original exact form.

6.1 Affine Projection Algorithm (APA)

The affine projection algorithm, first presented in [Ozeki & Umeda 84], was developed as a result of the efforts of overcoming the limitations of the classical NLMS for speech signals. The APA has properties that combine the advantages of the NLMS and the RLS algorithms. The NLMS algorithm is known for its low computational complexity. However, its convergence is slow and its tracking capability is poor for speech input. The RLS algorithm, on the other hand, has the same convergence for both coloured and white input signal, but its large computational load is a drawback. The affine projection algorithm has less computational complexity than the RLS algorithm, but much faster convergence than the NLMS algorithm for speech input signal. It actually was proposed as a generalization of the NLMS algorithm and is based on a multiple dimension projection per tap update [Gilloire 95].

The adjustment vector $\Delta\hat{\mathbf{h}}[n]$ of the filter coefficients has to satisfy p equations:

$$\mathbf{y}_p[n] = \mathbf{X}_p^T[n]\left(\hat{\mathbf{h}}[n] + \Delta\hat{\mathbf{h}}[n]\right) \tag{6.4}$$

or equivalently

$$\mathbf{X}_p^T[n]\,\Delta\hat{\mathbf{h}}[n] = \mathbf{y}_p[n] - \mathbf{X}_p^T[n]\,\hat{\mathbf{h}}[n] = \mathbf{e}_p[n] \tag{6.5}$$

where the projection order p is much smaller than the filter length L.
$\mathbf{X}_p[n]$ is a $(L \times p)$ matrix whose columns represent the p past input vectors of length L:

$$\mathbf{X}_p[n] = [\,\mathbf{x}_L[n],\ \mathbf{x}_L[n-1], \ldots \mathbf{x}_L[n-p+1]\,]\,. \tag{6.6}$$

$\mathbf{e}_p[n]$ and $\mathbf{y}_p[n]$ are vectors with the p past elements representing the error vector and the desired signal vector respectively.
From Eq. (6.5) $\Delta\hat{\mathbf{h}}[n]$ can be uniquely determined:

$$\Delta\hat{\mathbf{h}}[n] = \mathbf{X}_p[n]\ \left(\mathbf{X}_p^T[n]\,\mathbf{X}_p[n]\,\right)^{-1}\mathbf{e}_p[n] \tag{6.7}$$

Using the covariance matrix

$$\mathbf{R}_p[n] = \mathbf{X}_p^T[n]\,\mathbf{X}_p[n] \tag{6.8}$$

a decorrelation FIR filter vector $\mathbf{g}_p[n]$ can be defined

$$\mathbf{g}_p[n] = \mathbf{R}_p^{-1}[n]\,\mathbf{e}_p[n] \tag{6.9}$$

This vector filters the row vectors of $\mathbf{X}_p[n]$ in order to synthesize the adjustment vector $\Delta\hat{\mathbf{h}}[n]$. Thus, it can be written

$$\Delta\hat{\mathbf{h}}[n] = \mathbf{X}_p[n]\,\mathbf{g}_p[n] \tag{6.10}$$

The block diagram of the APA is presented in Figure 6.2.

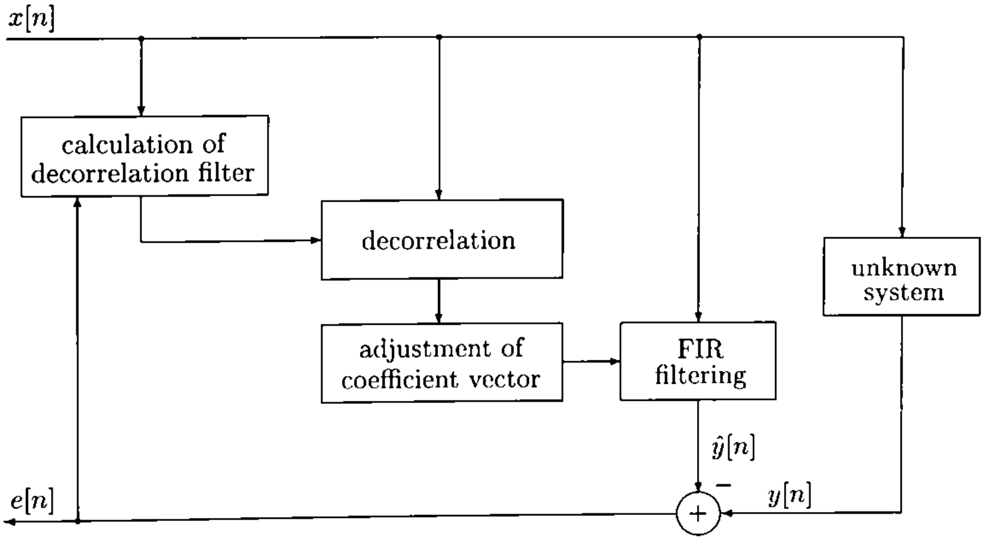


Figure 6.2: Block diagram of the conventional APA

Thus, the APA may be summarized as follows:

1. *filter output:*
$$\hat{\mathbf{y}}_p[n] = \mathbf{X}_p^T[n]\,\hat{\mathbf{h}}[n] \tag{6.11}$$

2. *estimation error:*
$$\mathbf{e}_p[n] = \mathbf{y}_p[n] - \hat{\mathbf{y}}_p[n] \tag{6.12}$$

3. *tap-weight adaptation:*
$$\hat{\mathbf{h}}[n+1] = \hat{\mathbf{h}}[n] + \mu\,\mathbf{X}_p[n]\,\left(\mathbf{X}_p^T[n]\,\mathbf{X}_p[n]\right)^{-1}\mathbf{e}_p[n] \tag{6.13}$$

$0 < \mu < 2$ being the stepsize that controls the amount of adjustment.
For the special case of $p = 1$ the above equations denote the NLMS algorithm and for $p = L$ they represent the RLS algorithm. Thus, the projection algorithm falls between the NLMS and the RLS algorithm.

The data matrix $\mathbf{X}_p[n]$ is a *Hankel* matrix, which has the property that all the elements along any cross diagonal are identical, i.e. $a[i,j] = a[i+j-n-1]$ with n denoting the number of columns of the matrix [Marple 87]. Hankel matrices can be related to Toeplitz matrices, so that the efficient inversion methods based on the Levinson algorithm can be applied[1]. As the product of two Hankel matrices is no longer Hankel, the simplified inversion method no longer applies to the term $\mathbf{X}_p^T[n]\,\mathbf{X}_p[n]$. This product of matrices can be inverted using the set of techniques known as *singular value decomposition* or *SVD* (Annex E).

With a speech input signal, the covariance matrix $\mathbf{X}_p^T[n]\,\mathbf{X}_p[n]$ may be ill conditioned[2], which results in large residual errors. To overcome this problem a regularization of the algorithm is performed:

$$\hat{\mathbf{h}}[n+1] = \hat{\mathbf{h}}[n] + \mu\,\mathbf{X}_p[n]\,\left(\mathbf{X}_p^T[n]\,\mathbf{X}_p[n] + \delta\mathbf{I}_p\right)^{-1}\mathbf{e}_p[n] \tag{6.14}$$

where $\delta \ll 1$ is the regularization parameter and $\mathbf{I}_p$ is the $(p \times p)$ identity matrix. By adding a small positive number to the diagonal of the covariance matrix, the term $\mathbf{X}_p^T[n]\,\mathbf{X}_p[n] + \delta\mathbf{I}_p$ will have δ as its smallest eigenvalue, even as $\mathbf{X}_p^T[n]\,\mathbf{X}_p[n]$ has eigenvalues close to zero. This yields a better conditioned inverse [Gilloire 95], provided that δ is large enough. It was found [Oh et al. 97] that the choice of the regularization parameter is very important in fixed-point implementation of the algorithm. In the regularization dominant case, when the covariance matrix is near zeros, after inversion the reciprocal value of δ will determine the maximum

[1] The Levinson algorithm is a recursive method of solving an L-th order symmetric Toeplitz system of equations:
$$\mathbf{A}\,\mathbf{x} = \mathbf{b}$$
where $\mathbf{A}$ is both symmetric and Toeplitz, and $\mathbf{x}$ and $\mathbf{b}$ are vectors [Stearns & David 93].

[2] A correlation matrix $\mathbf{R}$ is ill conditioned if the ratio of the largest eigenvalue to the smallest eigenvalue of $\mathbf{R}$ is large [Haykin 96].

values which can be represented in fixed-point. Thus, the larger δ, the less problems in the fixed-point implementation. Also, a too large value will degrade the performance, so that a compromise must be found.

Computational Load

The main cause of the computational burden of the APA is the inversion of the covariance matrix $\mathbf{R}_p^{-1}[n]$. The total computational complexity of the affine projection algorithm is about $(p+1)L+O(p^3)$ [Kaneda et al. 95], where $O(\cdot)$ denotes "order of" [Haykin 96]. Because of this complexity, the APA in its original form has been considered to be impractical.
To overcome this problem *fast* versions of the APA (FAP) were considered, which intend to reduce the computational requirements. There are several methods such as the recursive updating of the pre-filtering vector $\mathbf{g}_p[n]$, the update of an approximation filter instead of the estimation filter $\hat{\mathbf{h}}[n]$ [Tanaka et al. 95b] or the approximation of the covariance matrix $\mathbf{R}_p[n]$ for the inversion operation [Oh et al. 97]. This last approach forces a Toeplitz structure on the covariance matrix considering that for $p \ll L$ the following holds:

$$\hat{r}_\tau[n] \approx \hat{r}_\tau[n-1] \approx \ldots \approx \hat{r}_\tau[n-p+1] \tag{6.15}$$

Thus, the correlation matrix

$$\mathbf{R}_p = \begin{bmatrix} \hat{r}_0[n] & \hat{r}_1[n] & & \hat{r}_{p-1}[n] \\ \hat{r}_1[n-1] & \hat{r}_0[n-1] & & \hat{r}_{p-2}[n-1] \\ \hat{r}_2[n-2] & \hat{r}_1[n-2] & & \hat{r}_{p-3}[n-2] \\ \vdots & \vdots & & \vdots \\ \hat{r}_{p-1}[n-p+1] & \hat{r}_{p-2}[n-p+1] & \cdots & \hat{r}_0[n-p+1] \end{bmatrix} \tag{6.16}$$

can be approximated as:

$$\tilde{\mathbf{R}}_p = \begin{bmatrix} \hat{r}_0[n] & \hat{r}_1[n] & \cdots & \hat{r}_{p-1}[n] \\ \hat{r}_1[n] & \hat{r}_0[n] & \cdots & \hat{r}_{p-2}[n] \\ \hat{r}_2[n] & \hat{r}_1[n] & \cdots & \hat{r}_{p-3}[n] \\ \vdots & \vdots & & \vdots \\ \hat{r}_{p-1}[n] & \hat{r}_{p-2}[n] & & \hat{r}_0[n] \end{bmatrix} \tag{6.17}$$

which is a Toeplitz matrix. $\hat{r}_\tau[n]$ represents the estimate of the autocorrelation at lag τ and at time instant n, based on the last L input samples

$$\hat{r}_\tau[n] = \sum_{i=0}^{L-1} x[n-i]\,x[n-i-\tau]. \tag{6.18}$$

The inversion of $\tilde{\mathbf{R}}_p$ can be done efficiently. The updates for $\hat{r}_\tau[n]$ will be performed recursively by adding a new element corresponding to time instant n and

subtracting the element corresponding to time instant $n - L + 1$.
The fast version of the APA reduces the complexity of the conventional APA from $(p+1)L + O(p^3)$ to $2L + 20p$, where L represents the length of the estimation filter and p is the projection order. Thus it can be concluded, that the key features of fast affine projection algorithms include low complexity and memory requirements like the LMS and fast convergence for speech as input signal to the adaptation system like the RLS [Gay & Tavathia 95]. In Table 6.1 a comparison of the computational complexities of different adaptation algorithms is presented [Tanaka et al. 95a], [Haykin 96].

Adaptive algorithm	Complexity
NLMS	$2L$
RLS	$2L^2 + 8L$
Fast RLS	$8L$
Conventional APA	$(p+1)L + O(p^3)$
Fast APA	$2L + 20p$

Table 6.1: Comparison of computational complexity

The performance of the affine projection algorithm can be determined best by applying a white noise input signal which contains enough information for exciting all modes of the system. Thus, the convergence behaviour of the algorithm can be emphasized by using a white noise input signal because the ERLE (defined in Chapter 5, section 5.3.1) shows no fluctuations such as with normal speech input. Speech signals are characterized by sections of speech and pauses which, in the temporal course of the ERLE appear like increasing and decreasing segments. In Figure 6.3 the ERLE curves for a white noise input signal are pre-

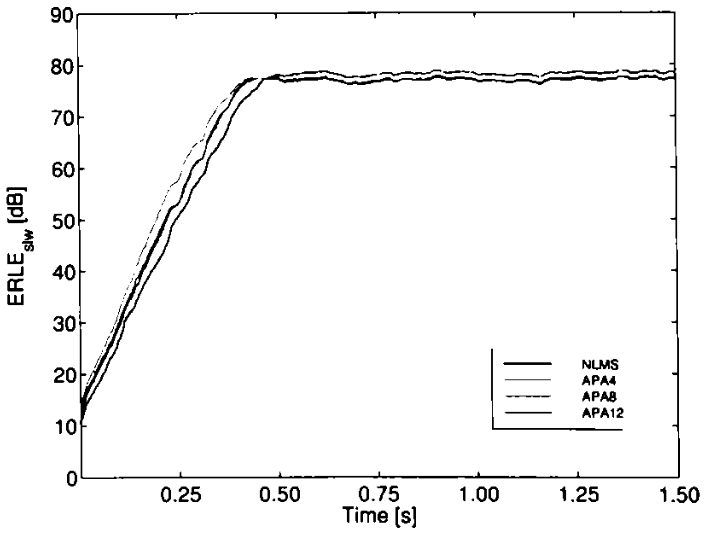


Figure 6.3: APA with white noise signal as input signal

sented. Different projection orders of the APA lead to different convergence time and, as expected, the higher the projection order, the faster the algorithm converges to the steady state. Furthermore it is apparent that the GSM convergence time requirement of 1 s for an ERLE of at least 20 dB are completely fullfilled.

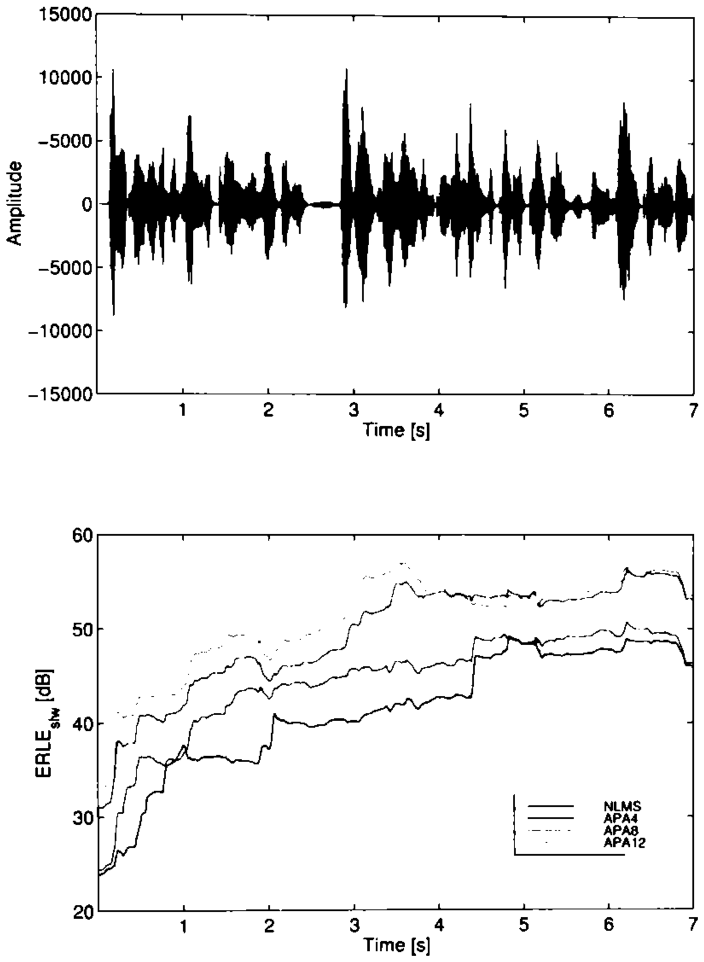


Figure 6.4: APA with speech input signal in noise-free environment

The affine projection algorithm, relying on a projection of the input signal vector, is much better suited for speech input signals than the widely used NLMS algorithm because of the decorrelation it performs on the input signal.

When estimating the performance of the APA, the objective measures presented in subsection 5.3.1 will be applied. The experiments carried out in this work start from the assumption that the length of the impulse response in a middle size car is usually around 30 ms. Applying a sampling rate of 8,000 Hz the transversal filter length has to be set to 256 coefficients for taking into account the whole length of the impulse response. The APA of projection order 1, 4, 8 and 12 were tested on different speech input signals in a noise-free near-end environment. The test signals were those presented in section 5.4.

Some experimental results are shown in Figure 6.4. They have been obtained by using the APA in its fast version, with the approximation of the covariance matrix by a Toeplitz structure as mentioned earlier in this section.
From Figure 6.4 it can be seen that in noiseless environments the convergence speed as well as the ERLE are increasing with increasing projection order of the APA. Thus, the NLMS, i.e. the first projection order of the APA, is the slowest from the convergence time point of view, while the APA of order 12 is the fastest. The best overall ERLE performance is obtained by the APA of dimension 12. However, after a certain interval of time, the performance of the APA of order 8 gets close to that of APA of dimension 12. The greatest difference between dimension 8 and 12 of the APA is notable during the first few seconds, when the ERLE performance of the dimension 12 APA is considerably better because of its shorter convergence time.

The microphone input signal of the AEC will additionally always be presented when the ERLE performance is discussed, because the temporal course of the performance measure is close related to this signal. Short speech pauses in the loudspeaker signal emitted into the car interior will lead to a falling tendency of the ERLE, because of the increasing error signal of the AEC during these periods. The special case of echo compensation in noisy environment will be considered later in this chapter in section 6.3.

6.2 APA in Subbands

The performance of the APA was also considered in the subband realization of the acoustic echo canceller. Subband adaptive filtering has the potential advantage of reducing computational complexity and improving convergence speed [Gilloire 95], [Gilloire & Vetterli 92]. The complexity reduction can be substantial as it is roughly proportional to the number of subbands. However, the use of nonideal critically sampled FIR multirate filterbanks leads to aliasing in the subbands which disturbs convergence and deteriorates the AEC performance.

A two channel filterbank consists of signal decomposition by highpass and lowpass half-band filters combined with a downsampling process. The data are reconstructed by upsampling and filtering as shown in Figure 6.5.
Aliasing cancellation at the synthesis bank output is achieved by using *Quadrature Mirror Filters (QMF)* with [Fliege 93]

$$\begin{aligned} H_0[z] &= H[z] \\ H_1[z] &= H_0[-z] \end{aligned} \qquad (6.19)$$

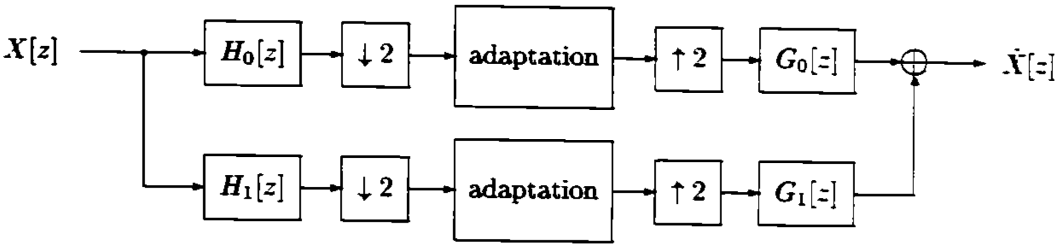


Figure 6.5: Two channel filter bank

$$\begin{aligned} G_0[z] &= 2H_0[z] \\ G_1[z] &= -2H_1[z] \end{aligned} \tag{6.20}$$

where $H[z]$ is a suitable lowpass filter prototype and $H[-z]$ the corresponding highpass filter. Under these conditions, the transfer function becomes allpass irrespective of the type and design methodology of the $H_0[z]$ and $H_1[z]$ half-band filters.

In Figure 6.6 the analysis and synthesis stages in polyphase structure are depicted. In the linear-phase FIR implementation of the polyphase structure, the half-band filter was designed by the standard Parks-McClellan method, the filter coefficients were taken from [Crochiere & Rabiner 83]. FIR filters lead to a signal delay depending on the number of filter coefficients used. However, the most disturbing effect for the AEC is the aliasing caused by flat filter slopes. To eliminate aliasing, a bandstop filter between the subbands was used.

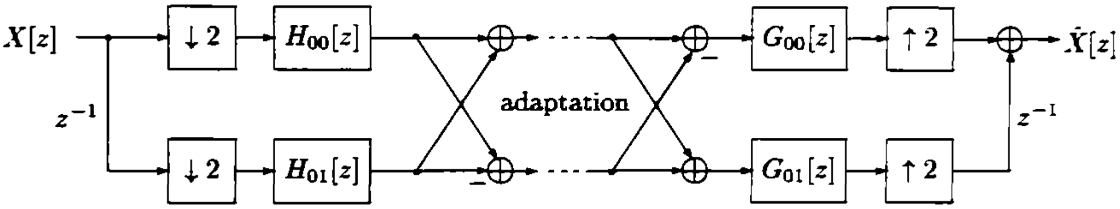


Figure 6.6: Analysis and synthesis stages of a QMF FIR filter bank in polyphase structure

Signal splitting in more than two subbands can be achieved by the tree structure [Vaidyanathan 93]. In Figure 6.7 the structure of an acoustic echo canceller in four subbands is presented, A standing for the analysis and S for the synthesis stages.

The APA performance in subbands was tested in a noise-free near-end environment on different speech input signal emitted by both male and female speakers. The average results of the mean ERLE values for different numbers of subbands and different orders of the affine projection algorithm are summarized in Table 6.2. It can be observed that for each APA dimension the $\text{ERLE}_{\text{mean}}$

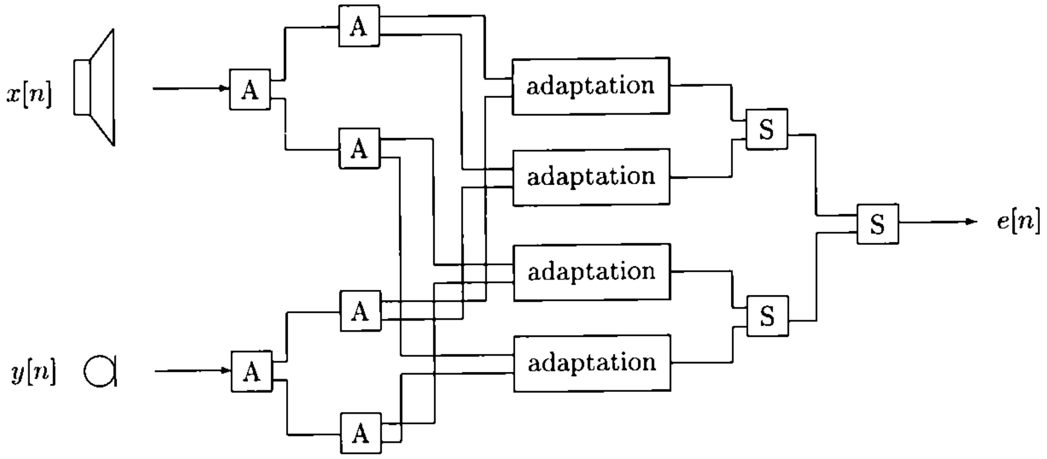


Figure 6.7: Subband AEC structure

gets higher with an increasing number of subbands. This is due to the additional decorrelation introduced by the filter bank.

However, if the performance in a certain filter bank structure is considered depending on the APA projection order, it can be observed that starting with a number of 8 subbands, the ERLE performance does not follow the increasing dimension of the affine projection algorithm. On the contrary, the mean ERLE even decreases for the APA of dimension 8 and 12. This is obviously a result of the decorrelation introduced by the filter bank combined with the decorrelation of the APA.

Regarding the projection order of the APA, it can be seen that the largest improvement of the ERLE is achieved when considering the transition from NLMS to APA of dimension 4. This is true for any subband implementation, the full-band to the 16 subbands approach. The increase in computational complexity for higher orders of the affine projection algorithm is not justified by any notable enhancement of the ERLE-gain.

As a conclusion to the results of Table 6.2 for a noise-free environment on the near-end side, it can be suggested to limit the projection order of the adaptation algorithm to 4. For this APA dimension, every increase in the number of subbands of the filter bank is reflected in a better ERLE performance. Higher orders of the affine projection algorithm do not present a steadily increasing ERLE performance with an increasing number of subbands. Thus, for the 16 subbands realization of the filter bank, the ERLE results are decreasing for increasing projection orders of the APA.

In Figure 6.8 the temporal course of the ERLE for a speech input signal ap-

Dimension of APA	Number of Subbands	$\text{ERLE}_{\text{mean}}$ [dB]
1	1	+14.54
1	4	+20.83
1	8	+32.40
1	16	+43.12
4	1	+18.39
4	4	+26.57
4	8	+37.52
4	16	+46.23
8	1	+22.46
8	4	+29.34
8	8	+38.76
8	16	+45.09
12	1	+22.81
12	4	+30.09
12	8	+38.00
12	16	+43.04

Table 6.2: $\text{ERLE}_{\text{mean}}$ for a noise-free near-end environment

plied to an affine projection algorithm of dimension 4 is presented. It can be seen that the performance of the ERLE and the convergence time are getting better with increasing number of subbands in the filter bank.

Figure 6.9 shows a comparison between the temporal course of the ERLE for the projection order 1 (NLMS) and 12 of the adaptation algorithm, in different subband structure implementations (fullband, 4 subbands, 8 subbands and 16 subbands). It can be observed that for the filter bank with 16 subbands there is almost no difference in ERLE performance when comparing the results for the NLMS and the APA of order 12. The ERLE performance difference increases with a decreasing number of subbands. Thus, in the fullband and 4 subbands realization the most notable enhancement in the performance of the APA with dimension 12, compared to that of the NLMS, is registered. It is also worth mentioning that the ERLE curves get flatter when more subbands are used, this being a result of the decorrelation of the speech input signal performed by the filter bank.

In [Ansahl et al. 98] it is found that the use of IIR half-band filters and biquadratic notch filters is to be preferred to the FIR implementation of the filterbank with antialiasing bandstop filters. The ERLE performance is nearly the same, the advantages lie in the reduction of computational complexity and

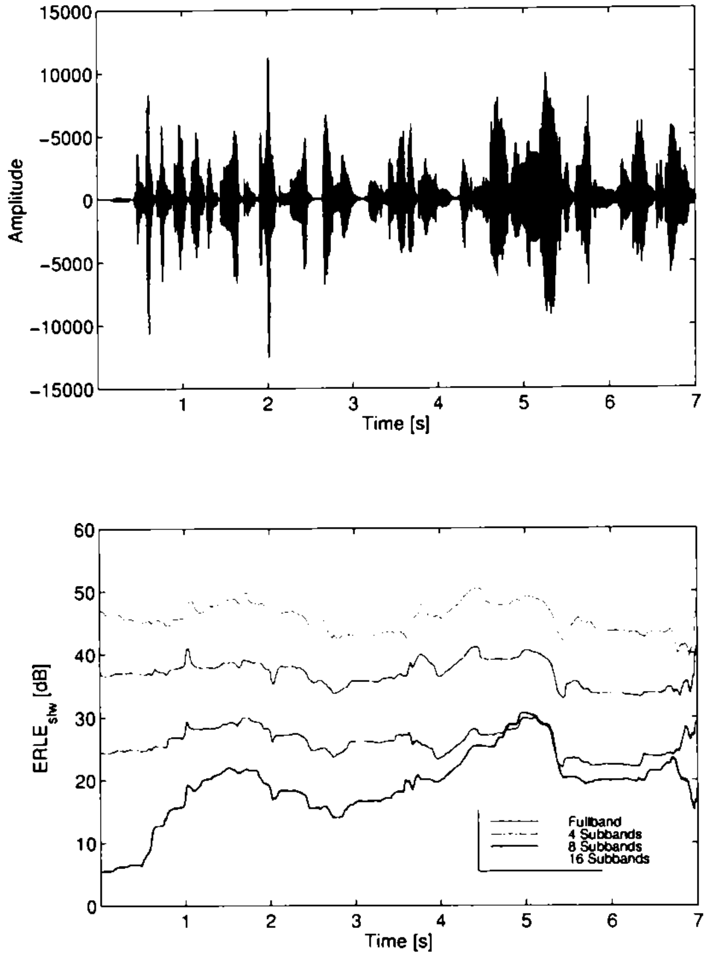


Figure 6.8: Subband APA 4 performance for speech signal in noise-free environment

the sharp frequency separation of the IIR filters.

6.3 APA in Noisy Environment

In the following an original derivation of the influence of a noisy near-end environment on the performance of the affine projection algorithm will be presented [Kremmer & Ansahl 98]. Assuming the presence of noise at the near-end, the desired signal will be

$$\mathbf{y}'_p[n] = \mathbf{y}_p[n] + \mathbf{n}_p[n] \tag{6.21}$$

where $\mathbf{n}_p[n]$ represents the noise vector containing the past p noise samples. The noisy desired signal $\mathbf{y}'_p[n]$ leads to an estimation error $\mathbf{e}'_p[n]$ which will also be noise-dependent

$$\mathbf{e}'_p[n] = \mathbf{y}'_p[n] - \hat{\mathbf{y}}_p[n] \tag{6.22}$$

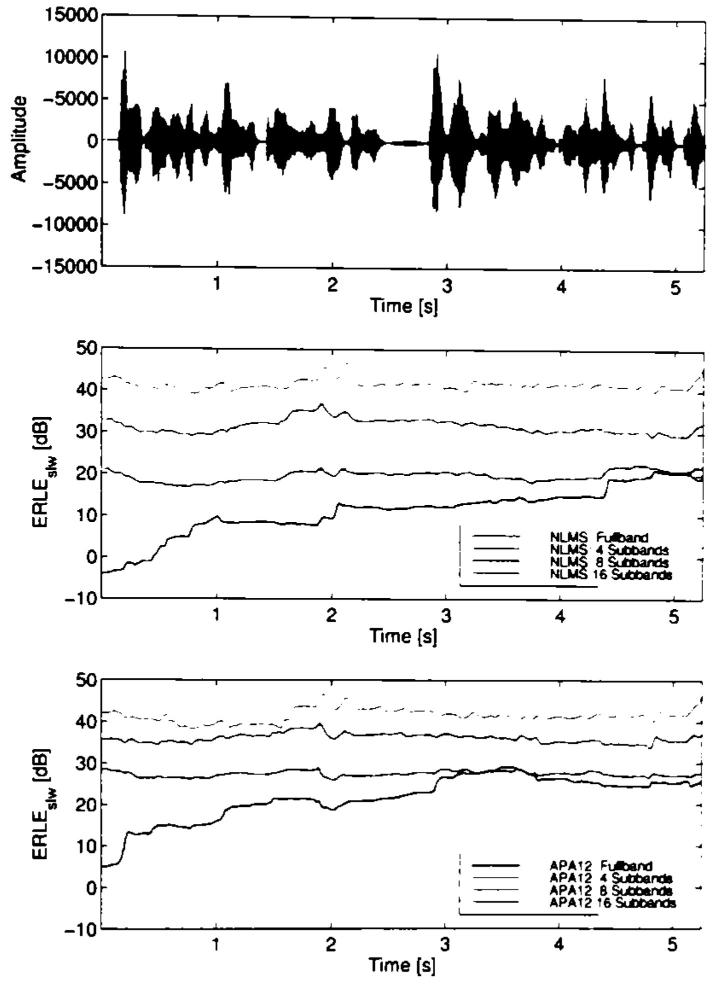


Figure 6.9: Comparison of subband APA performance in noise-free environment

Substituting Eq. (6.21) in Eq. (6.22), the error estimate in noisy environment will represent the sum of the error estimate in noise-free environment and the noisy near-end signal:

$$\mathbf{e}'_p[n] = \mathbf{e}_p[n] + \mathbf{n}_p[n] \tag{6.23}$$

Starting with the tap-weight adaptation equation (6.13) and considering Eq. (6.23), the tap-update equation for a noisy near-end environment will change to

$$\begin{aligned}\hat{\mathbf{h}}[n+1] &= \hat{\mathbf{h}}[n] + \mu\,\mathbf{X}_p[n]\left(\mathbf{X}_p^T[n]\,\mathbf{X}_p[n]\right)^{-1}\mathbf{e}_p[n] \\ &\quad + \mu\,\mathbf{X}_p[n]\left(\mathbf{X}_p^T[n]\,\mathbf{X}_p[n]\right)^{-1}\mathbf{n}_p[n] \end{aligned} \tag{6.24}$$

Eq. (6.24) shows the tap-weight adaptation and its dependency on the near-end noise signal. It can be seen that the disturbing effect of the near-end noise will be reinforced with higher projection orders of the APA.

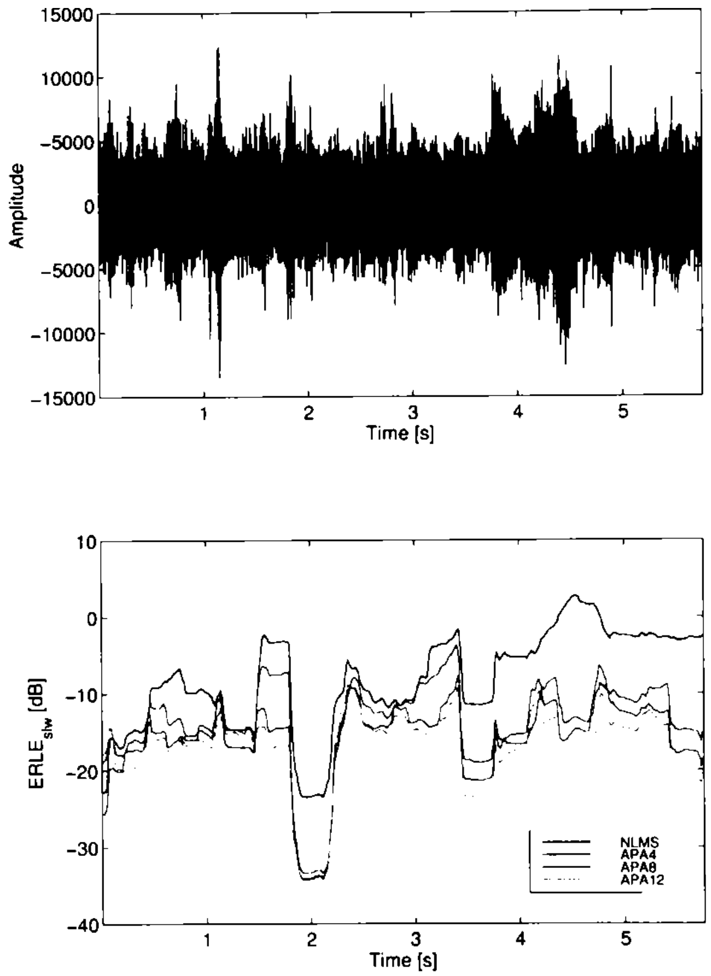


Figure 6.10: Fullband APA performance with speech input signal, noisy environment of SNR = −10 dB

6.3.1 In the Fullband

As a consequence of Eq. (6.24), when noise is present at the microphone input an inversion of the ERLE results obtained with clean echo signals has to be expected, namely the lower the projection order of the algorithm the better the anticipated ERLE. According to Eq. (6.24), the higher the projection order the longer the APA will stay under the influence of a noise event, because in its data matrix more past input vectors are considered and compared to the noisy desired signal vector. Thus it can be assumed to get a better ERLE result for the NLMS than for an APA of higher dimension.

The theoretical conclusions from Eq. (6.24) can be confirmed by the investigations made on noisy near-end signals. The APA was considered in different projection orders, and different noisy environments were examined. The result-

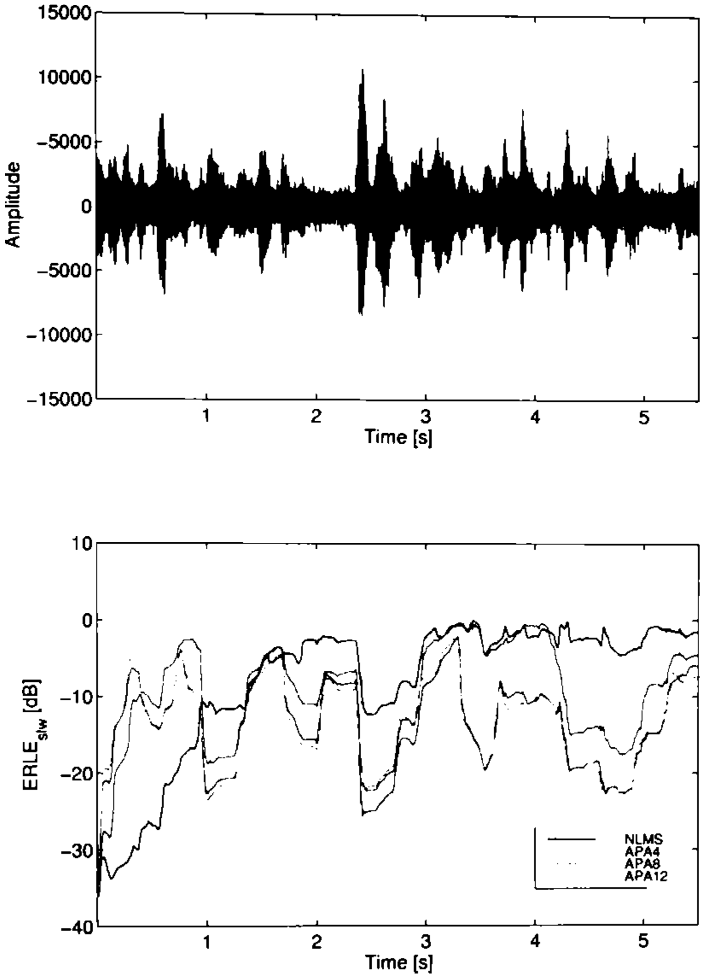


Figure 6.11: Fullband APA performance with speech input signal, noisy environment of SNR = 0 dB

ing ERLE curves for near-end noisy environments of SNR of −10 dB and 0 dB are presented in Figures 6.10 and 6.11. In both, the overall better performance of the NLMS can be viewed. At the beginning, an initial period can be recognized, where the convergence performance of the APA in dimension 12 is better than that of the other dimensions of the affine projection algorithm. The noisier the near-end environment the faster the NLMS will get to its best performance. Thus, the ERLE curve corresponding to the NLMS algorithm will achieve its greatest enhancement much faster in the noisy environment of −10 dB than in that of 0 dB, compared to the other dimensions of the APA.

The average results of the ERLE$_{\mathrm{mean}}$ values for different speech input signals in different numbers of subbands and for different orders of the affine projection algorithm are summarized in Table 6.3. For the implementation in the fullband, in noisy near-end environment of different SNRs, it can be observed that the ERLE performance decreases with increasing dimension of the APA. This shows the correspondence between experimental results and the conclusion from the theoretical derivation of section 6.3.

6.3.2 In Subbands

In this section the performance of the affine projection algorithm in the subband implementation will be investigated in noisy near-end environments. This study has also been presented in [Kremmer & Ansahl 98]. The results obtained in this investigation are presented in Table 6.3. It can be observed that with increasing number of subbands the ERLE performance is being enhanced. The decomposition of the loudspeaker speech input and of the noisy microphone signal combined with the adaptation performed in the subbands, leads to an increase of ERLE performance with increasing number of subbands. The special case of fullband implementation, as presented in the previous section, is excluded from this statement. When analyzing these results, it can be observed that for each APA dimension the mean ERLE value raises with increasing number of subbands, as in the case of noise-free near-end environment.

Number of Subbands	Dimension of APA	ERLE$_{\text{mean}}$ [dB] SNR = −10 dB	ERLE$_{\text{mean}}$ [dB] SNR = 0 dB	ERLE$_{\text{mean}}$ [dB] SNR = +10 dB
1	1	− 9.39	− 2.67	+ 6.24
1	4	−10.32	− 3.60	+ 5.30
1	8	−11.44	− 4.72	+ 4.18
1	12	−12.81	− 6.09	+ 2.82
4	1	− 6.42	+ 0.25	+ 8.60
4	4	+ 4.07	+10.66	+18.56
4	8	+ 5.24	+11.88	+20.12
4	12	+ 4.01	+10.68	+19.18
8	1	− 0.51	+ 6.22	+15.05
8	4	+11.57	+18.24	+26.78
8	8	+10.62	+17.32	+26.01
8	12	+10.22	+16.91	+25.59
16	1	+ 5.95	+12.66	+21.53
16	4	+14.90	+21.61	+30.40
16	8	+13.86	+20.57	+29.37
16	12	+13.92	+20.61	+29.34

Table 6.3: ERLE$_{\text{mean}}$ in different subband structures and for different noise conditions

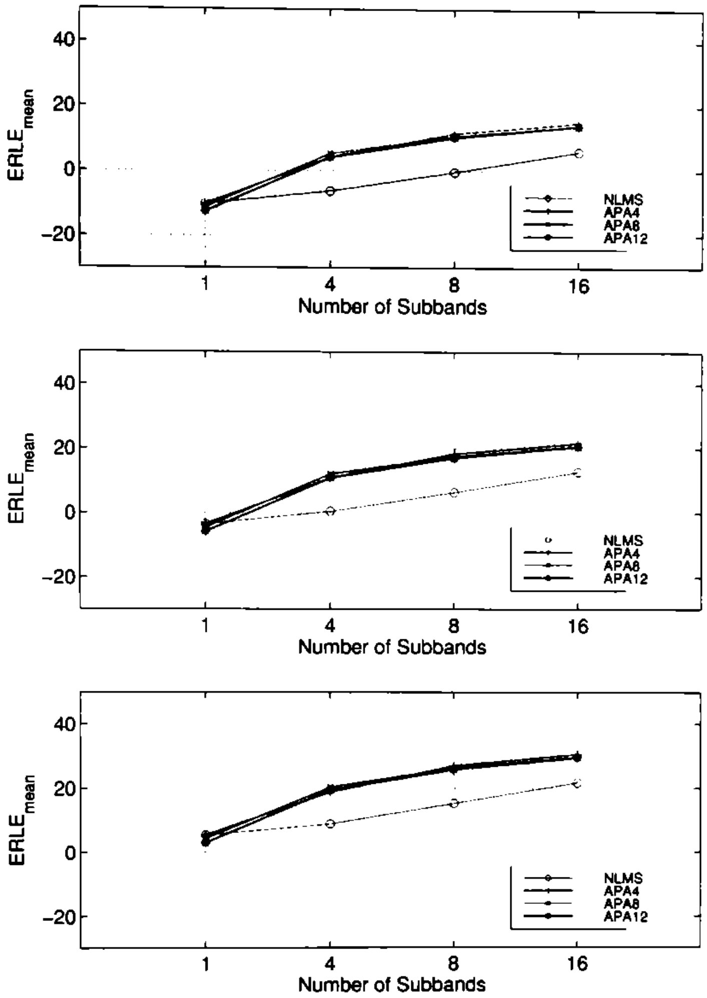


Figure 6.12: ERLE$_{\text{mean}}$ performance in subbands for different APA dimensions in noisy environments of -10 dB, 0 dB and $+10$ dB

If the ERLE$_{\text{mean}}$ performance in a specific filter bank structure is considered, it can be observed that the falling tendency of the mean ERLE value, also noticed in the noise-free near-end environment investigations presented in Table 6.2, already starts in the 4 subbands approach. However, this holds only for APA of projection dimension 12. In the 4 subbands implementation an continuous mean ERLE enhancement can be confirmed only for the projection orders 1, 4 and 8.

Starting from a number of 8 subbands, the ERLE does not increase any more for the APA of dimension orders 8 and 12, on the contrary, it decreases. This is obviously a result of the interaction of the decorrelations introduced both by the filter bank and the affine projection algorithm.

Regarding the preferred dimension of the APA, it can be seen that the most

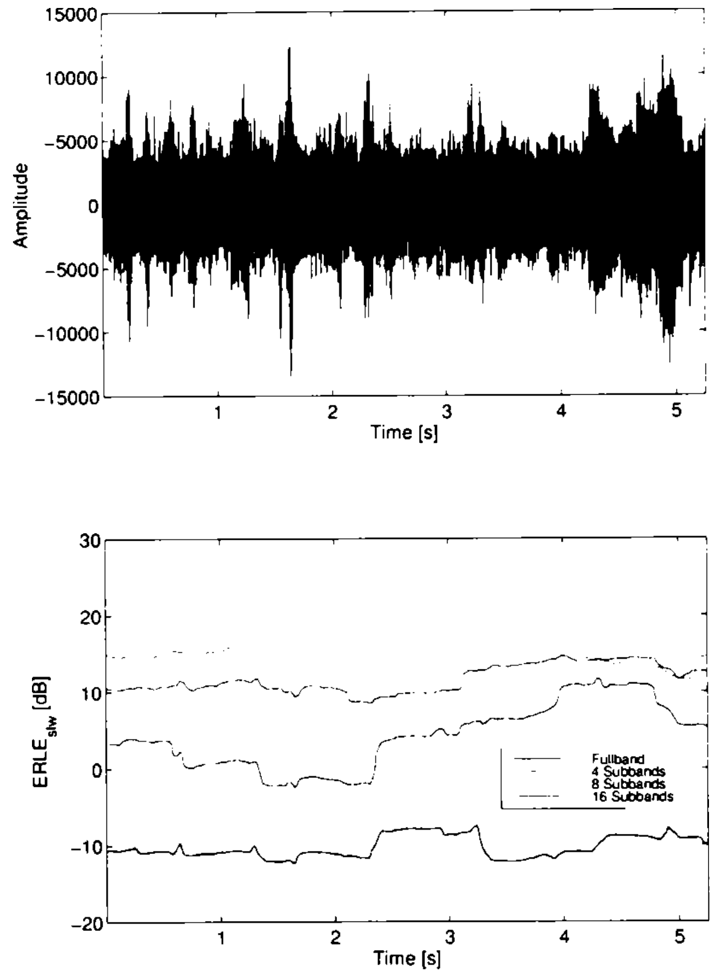


Figure 6.13: Subband APA 4 performance for speech signal in noisy environment (SNR = −10 dB)

pronounced refinement on the echo compensation task is registered at the transition from the adaptation with NLMS to the APA of order 4, just as in the noise-free case. This is true for any subband implementation, from the fullband to the 16 subbands approach.

As a conclusion to the subband approach results of Table 6.3 for a noisy near-end environment, the suggestion from the noise-free discussion can be validated. Thus, the higher the filter bank order the better are the results for an APA of dimension 4, both in noisy and in noise-free near-end environments. Because there is plenty of information contained in Table 6.3, it is helpful to visualize the results in a more accessible way in Figure 6.12. The three diagrams show the dependency of the mean ERLE on the number of subbands in the filter bank for the noisy environments of −10, 0 and 10 dB.

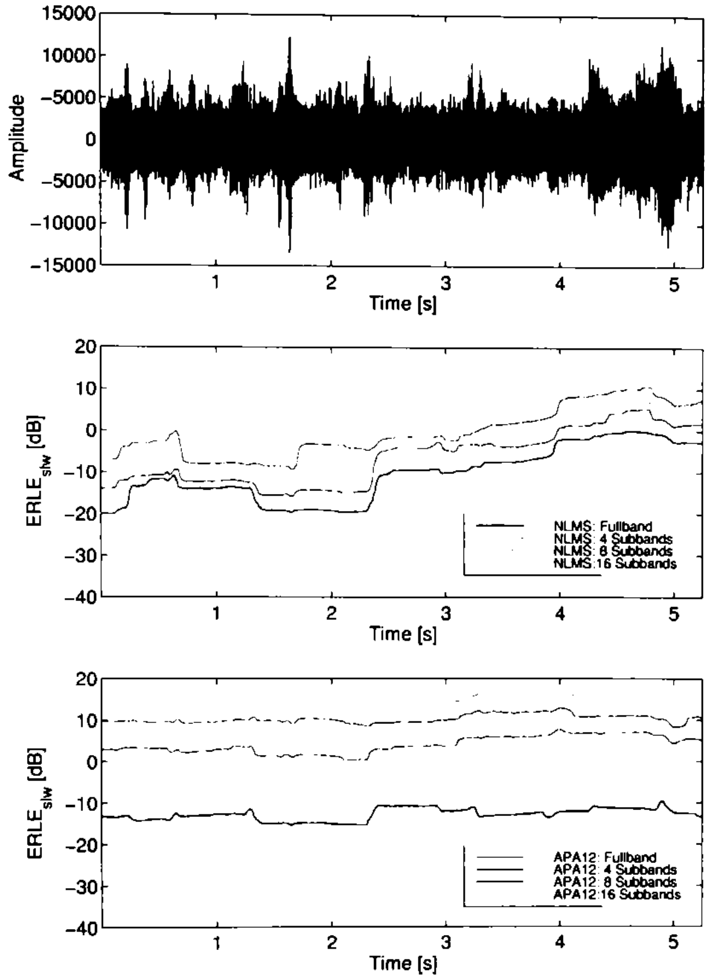


Figure 6.14: Comparison of subband APA performance in noisy environment (SNR = −10 dB)

In Figure 6.13 the ERLE temporal course for a speech input signal applied to an affine projection algorithm of dimension 4 is presented in a noisy environment of about −10 dB. It is obvious that the performance of the ERLE is increasing with the number of the subbands in the filter bank. Furthermore, a flattening of the ERLE curves for higher filter bank orders can be observed.

Figure 6.14 shows the temporal course of the ERLE for the projection order 1 and 12 of the adaptation algorithm in different subband structures (fullband, 4 subbands, 8 subbands and 16 subbands) for a noisy near-end environment of about −10 dB. A general enhancement can be registered for each subband realization. Thus, unlike the noise-free case (Figure 6.9), the filter bank with 16 subbands still shows a performance enhancement when comparing the ERLE curves of the NLMS and the APA of dimension 12.

It is also worth mentioning that the ERLE curves get flatter when higher orders of the affine projection algorithm and more subbands are used, being a result of the decorrelation of the speech input signal performed by the adaptation algorithm and the filter bank.

6.4 Stepsize Control of the APA in Noisy Environment

To make the acoustic echo cancellation algorithm noise robust, a stepsize control was developed which varies the adaptation coefficient value depending on the ratio of the loudspeaker signal power to the estimated near-end noise power. The algorithm, the test results and the resulting conclusions are also presented in [Kremmer 98]. The algorithm takes into account the nonstationarities of the input signal as well as the noise at the near-end.
In [Meana et al. 94] and [Hirano & Sugiyama 95] it was shown for the NLMS algorithm that the noise influence at the near-end becomes very important for small input signal power and that the NLMS algorithm cannot update its filter coefficients correctly any more. It is advisable to select a small adaptation coefficient when operating in high background noise, but at the same time, the adaptation should converge rapidly to the true echo path, which implies the use of a large stepsize.

The proposed adaptation algorithm can be described as follows:

$$\mu[i] = \frac{\mu_0}{1 + \alpha \dfrac{P_x[i]}{P_n[i]}} \tag{6.25}$$

where $\mu[i]$ represents the variable stepsize for frame i and μ_0 is the constant stepsize. $P_x[i]$ and $P_n[i]$ represent the average power of the loudspeaker signal and the near-end noise power for frame i, respectively. The near-end noise power will be estimated by a VAD and updated during periods of speech-free frames. α is a weighting factor. The stepsize will be updated on frame basis, once every 15 to 30 ms. During these segments, the speech and noise power will be averaged.

After a series of computer simulations with different near-end noise and loudspeaker signals, optimal values for μ_0 and α were found. Good results were obtained when μ_0 was varied from 0.5 to 0.9. For the weighting factor α a logarithmic dependency on the ratio $\frac{P_x[i]}{P_n[i]}$ was experienced. Thus, for different power ratios different weighting factors will be taken. In the implemented version, this dependency was approximated by 5 different values for α.

Noise level	$\Delta\mathrm{ERLE}_{\mathrm{mean}}$	$\Delta\mathrm{ERLE}_{\mathrm{mean}}$	$\Delta\mathrm{ERLE}_{\mathrm{mean}}$	$\Delta\mathrm{ERLE}_{\mathrm{mean}}$	Noise
[dB]	[dB]	[dB]	[dB]	[dB]	type
	NLMS	APA 4	APA 8	APA 12	
−10	16.98	11.55	7.12	5.64	car noise
0	11.00	6.20	3.23	2.34	car noise
+10	5.24	1.43	0.22	0.13	car noise
−10	8.34	3.52	2.19	1.69	white noise
0	5.17	1.42	0.66	0.48	white noise
+10	2.66	0.52	0.28	0.20	white noise

Table 6.4: Enhancement of $\mathrm{ERLE}_{\mathrm{mean}}$ using a variable stepsize

The $\mathrm{ERLE}_{\mathrm{mean}}$ enhancement is defined as being the difference between the mean ERLE value achieved by using the stepsize control and the mean ERLE value without using the stepsize control function. Table 6.4 summarizes the results of the $\mathrm{ERLE}_{\mathrm{mean}}$ enhancement in different noisy environments. Different SNRs and noise types are considered. Analyzing the results, it can be observed that the improvement of the mean ERLE performance depends on:

- the projection order of the affine projection algorithm, the best enhancement being achieved by the NLMS, the smallest by the APA with projection order 12

- the signal-to-noise ratio at the near-end, the best results being obtained in an noisy environment of about −10 dB

- the nature of the background noise, better results were obtained for car noise than for white noise.

The temporal course of the ERLE for two different noise conditions and dimensions of the APA are presented in Figures 6.15 and 6.16. The sliding window ERLE curves for the NLMS algorithm in a noisy environment of about −10 dB with and without the proposed stepsize control function is presented in Figure 6.15. Considering the stepsize variation curve, it can be observed that the change of the variable stepsize value $\mu[i]$ towards lower values always occurs when the clean loudspeaker signal has small values and therefore the noise influence at the near-end becomes very large. A reduction of the stepsize means a slower convergence of the adaptation algorithm. Consequently, the adaptation will not go into the wrong direction for small loudspeaker input values for adverse noise conditions at the near-end.
Thus, the stepsize control function leads to a much better adaptation compared to the fix-valued stepsize approach.

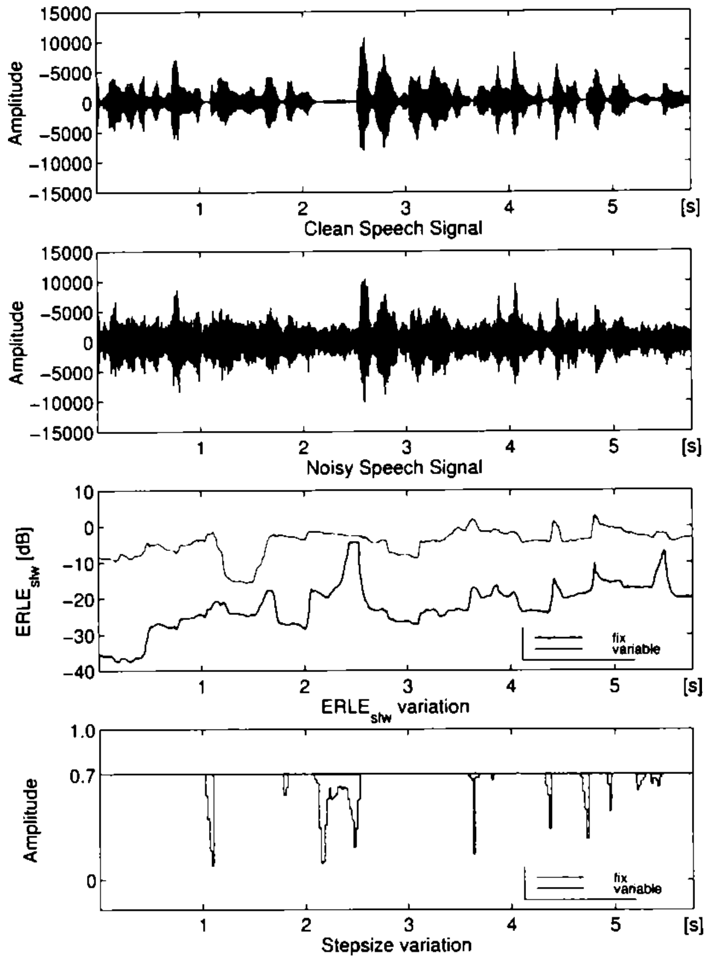


Figure 6.15: NLMS with stepsize control in noisy environment (SNR = −10 dB)

As already stated when analyzing the results of Table 6.4, the improvement of the ERLE gets smaller for better noise conditions at the near-end and for higher orders of the adaptation algorithm. This can be viewed in Figure 6.16 which presents the case of an affine projection algorithm with dimension 8 in a noisy environment of about 0 dB SNR. Here the segments of variable stepsize value are less and shorter in duration than in the case presented in Figure 6.15 for a noisier near-end environment.

6.5 Conclusions

Based upon the decision made in Chapter 4 considering the optimization criterion for the filter design of the acoustic echo canceller, the affine projection algorithm (APA) was chosen to be applied in the proposed combined system.

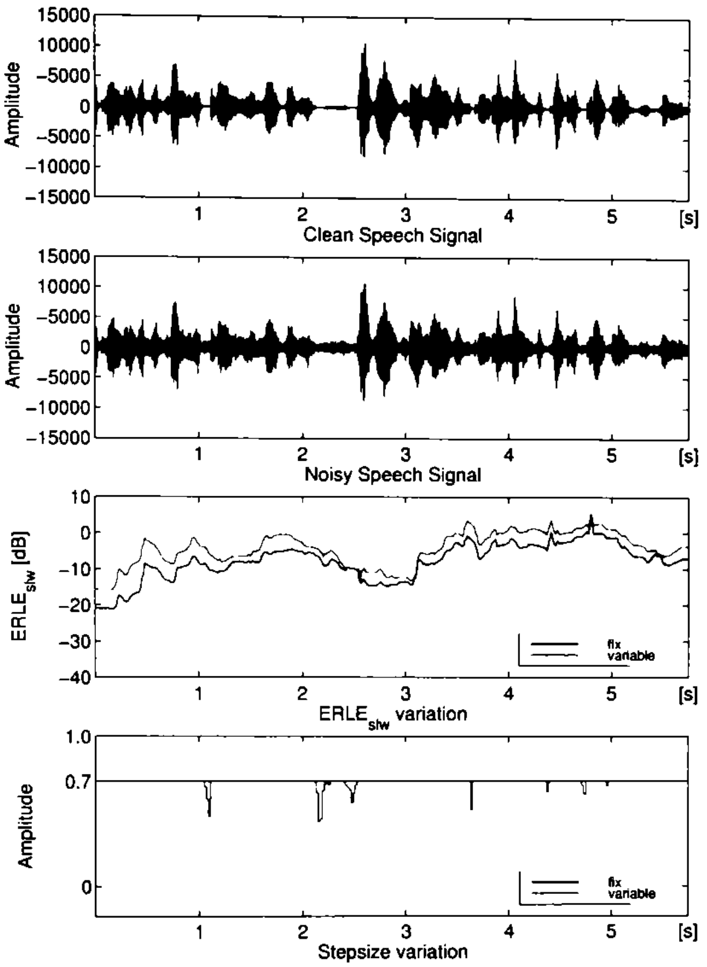


Figure 6.16: APA 8 with stepsize control in noisy environment (SNR = 0 dB)

The APA has properties that combine the advantages of the NLMS and the RLS algorithms. The affine projection algorithm has less computational complexity than the RLS algorithm, but much faster convergence than the NLMS algorithm for a speech input signal. It actually was intended to be a generalization of the NLMS algorithm and it is based on a multiple dimension projection per tap update. The computational burden of the original affine projection algorithm is determined by a matrix inversion that has to be performed. The fast affine projection algorithm has reduced computational requirements because of the recursive updating it makes use of.

The performance of the APA has been investigated in the fullband and subbands approach. The filter bank implementation seemed to be worth examining because of the complexity reduction it promises. Thus, it is known that subband adaptive filtering, besides convergence improvements, also has the potential ad-

vantage of a computational reduction roughly proportional to the number of subbands. After a short presentation of filter bank theory, the APA performance was tested in the fullband, in 4, 8 and 16 subbands.
When a subband implementation of the AEC is intended, on the basis of the achieved results, it can be suggested to limit the dimension of the affine projection algorithm to 4 for a noise-free near-end environment. For the APA of dimension 4 every increase of the number of subbands in the filter bank is reflected in a better ERLE performance. Higher dimensions of the affine projection algorithm lead to worse results for higher filter bank orders compared to the performance of the adaptation with projection order 4.

Acoustic echo cancellation usually has to perform in noisy environments. Therefore an original derivation of the influence of a noisy near-end environment on the performance of the affine projection algorithm has been presented and the theoretical results have been confirmed by experimental results. Thus, it can be stated that the disturbing effect of a noisy environment is strengthened with increasing projection order of the APA. This can be explained in the following way: the higher the projection order the longer the APA will stay under the influence of a noise event, because in its data matrix more past input vectors are considered and compared to the noisy desired signal vector. Thus it is expected and confirmed that better ERLE results are obtained for the NLMS than for an APA of higher dimension.
However, if the performance in certain filter bank structures is considered, it can be observed that starting with a number of 8 subbands, the ERLE performance does not follow the increasing dimension of the affine projection algorithm. On the contrary, the mean ERLE even decreases for the APA of dimension 8 and 12.

The ERLE performance in the subband structure in noisy near-end environments has been investigated and also presented in [Kremmer & Ansahl 98]. It has been found that as in the case of noise-free near-end environment, for each APA dimension the mean ERLE raises with increasing number of subbands. If the ERLE performance in a specific filter bank structure is considered, it can be observed that the falling tendency of the mean ERLE value, also noticed in the noise-free near-end environment investigations, already starts in the 4 subbands approach. However, this holds only for the APA of dimension 12. In the 4 subbands implementation a steady enhancement of the $\mathrm{ERLE}_{\mathrm{mean}}$ can be confirmed only for the projection orders 1, 4 and 8.
Starting from a filter bank with 8 subbands, the ERLE does not increase any more for APA dimensions 8 and 12, on the contrary, it decreases. This is obviously a result of the interaction of the decorrelations introduced both by the filter bank and the affine projection algorithm.

For the fullband approach in noisy environment, a new stepsize control algorithm has been proposed, implemented and tested. The algorithm varies the usually fixed adaptation coefficient μ in accordance to the ratio of the loudspeaker signal power and the estimated noise power at the microphone input. The algorithm, also presented in [Kremmer 98], shows best ERLE improvements for the NLMS in very noisy environment. An increase in the APA dimension and better noise conditions at the near-end lead to smaller ERLE improvements.

Summarizing the results for the subband approach in noiseless and noisy near-end environments, it can be suggested to choose an affine projection algorithm of dimension 4. The higher the selected filter order the better the ERLE and convergence performance. Dimension orders of the affine projection algorithm greater than 4 do not justify the increase in computational complexity.

For the fullband approach in noise-free environment the APA of projection order 8 can be recommended, higher orders of the affine projection algorithm do not justify the increase in computational complexity. When operating in adverse near-end conditions the use of NLMS combined with an adaptive stepsize control function is advisable. Thus, as a comprimise, working in both noisefree and noisy environments, the APA of projection order 4 with adaptive stepsize control will be preferred.

Chapter 7

Speech Detectors

In acoustic echo cancellation as well as in noise reduction systems, there is need of an algorithm that has the ability to decide whether a signal segment contains speech or only noise. When no far-end speech is present the AEC has to stop its operation and in the case of speech enhancement the estimation of the background noise will be performed during speech pauses.
It is known that approximately 20% of normal speech consists of pauses, which occur anywhere between spoken words and sentences [Armbruster et al. 91]. The process of distinguishing between speech and nonspeech sections in a speech signal is called *voice activity detection.*
The function of a voice activity detector in noisy environment is to differentiate between speech superimposed on the background noise and speech-free noise. The performance of the algorithm is a function of both the noise level (SNR) and the structure of the noise (stationary, nonstationary, white or periodic) [El-Maleh & Kabal 97] and is characterized by the degree and severity of speech clipping and the percentage of speech activity[1] it indicates. If the VAD fails to detect every speech event, speech quality will be degraded by clipping.

The decision of the VAD should be "fail-safe", i.e. if the decision is in doubt, it should indicate "speech present", because it is more harmful to classify speech as noise as the other way round. Another requirement on the VAD algorithm is the possibility to self-adapt to the changing level of background noise, so that its relevant thresholds should be determined from measurements made directly on the processed input signal [Rabiner & Sambur 75]. To give reliable detection, the threshold must be sufficiently above the noise level, otherwise noise could be identified as speech, but not so far above as to miss low level parts of speech by interpreting them as noise.

During recognition of speech segments, the VAD must present a *fast attack*

[1] The percentage of speech activity is the percentage of time during which the VAD is active.

period, which means that the algorithm must identify precisely the beginning of an utterance. This is a difficult task especially when an utterance begins with weak, i.e. low energy fricatives (/f/, /th/, /h/) or weak plosive bursts (/p/, /t/, /k/) [Rabiner & Schafer 78]. At the end of an utterance the algorithm has to declare the first few frames of silence after a detected speech burst to still be speech. This procedure is called *hangover* and it minimizes the probability of missing low-energy unvoiced speech at the end of the utterance, such as final nasals, voiced fricatives (/v/), which become devoiced at the end of words or trailing off of vowel sounds.

The VAD hangover period is very important in eliminating mid-burst clipping of low level speech [GSM Rec. 06.32 95]. There is a certain minimum duration a speech burst must exceed before it is prolonged by adding the hangover. Otherwise, noise spikes, falsely detected as speech, could be extended. The hangover mechanism is not efficient in correcting isolated VAD errors, e.g. a 1 among a sequence of zeros or vice versa. Such errors can be corrected by accepting a delay of 2 or 3 frames in the VAD decision and monitor the decisions in neighbouring frames. If the VAD decision of the current frame is different from that of the close neighbours, the VAD flag of the current frame is changed to be similar to the decision of the neighbouring frames [El-Maleh & Kabal 97]. This procedure is repeated for every frame, to remove any isolated errors.

7.1 Voice Activity Detection

Accuracy, robustness to noise, simplicity, adaptation and real-time processing are some of the required features of a good VAD.
The basic principle of a VAD is that it extracts some measured features or quantities from the input signal and then compares these values with thresholds, usually extracted from speech-free periods [El-Maleh & Kabal 97]. The design of a VAD consists in selecting these features for the speech/noise decision, and the definition and update rules for the thresholds. If the measured values exceed the thresholds, voice activity is assumed. The VAD algorithm outputs a binary decision on a frame-by-frame basis, where a frame is usually 20-40 ms long.

The most common features used in the detection process of speech in noise are

- short-time energy
- zero crossing rate
- LPC coefficients.

More complex VADs use cepstral features, formant shape [Hoyt & Wechsler 94] or least-squares periodicity measures [Tucker 92] as decision features. Concave or convex formant patterns could be observed in speech but not in noise. Thus, analyzing the formant shape could give information about the presence of speech in noise.

The simplest approach to speech detection is an energy detector, which compares the short-term energy of the input signal to its long-term energy or to a predefined energy threshold. An energy based algorithm which makes use of two thresholds is presented in [Harrison et al. 86]: the lower threshold is set to 1.2 times the estimated background noise energy and the upper to 1.5 times the noise energy. The beginning point of speech is chosen as the point when the signal energy last crosses the lower threshold, before it crosses the upper threshold. Correspondingly, the ending point of speech is considered to be the point when the signal energy first crosses the lower threshold after it has crossed the upper threshold. The algorithm is a sample-by-sample realization and thus computationally very demanding.
The energy computation can be performed in the frequency domain as well [Pollak et al. 93]. In this case the algorithm works on signal segments rather than for each sample of the signal, the segment energy being calculated from the DFT coefficients of the considered signal segment. The algorithm is simple, but gives good results only for positive SNRs.

Using adaptive thresholds which follow the changing level of estimated background noise, a certain initial estimate of the noise parameters must be considered. Usually the assumption is made that the first period of the input signal is speech-free. This period is appreciated to be in the range of 100 ms [Rabiner & Schafer 78] over 320 ms [ITU Rec. G.729] to even 1 second as it is considered to be necessary in [Harrison et al. 86].

Another possibility of detection in the frequency domain consists in spectra comparison, where low energy unvoiced sounds of high frequency and high energy voiced sounds of low frequency equally contribute to the result. This makes it easier to detect weaker high frequency sounds. A subband approach for emphasizing the contribution of the unvoiced signal is considered in [Yang 93]. The SNR factors in different subbands will be evaluated. Since the high frequency components of background noise are relatively small, the strong high frequency unvoiced sound can be detected. The voice detection criterion used in this approach is the mean of the SNR factors of the different subbands, which will be compared to a predetermined SNR threshold. The SNR_k for the k-th subband, is defined as the maximum positive SNR out of the SNRs corresponding to the different frequency bins within the respective subband. In the calculation of the SNR_k factors the continuously updated noise power estimate is taken into account.

Considering the background noise to be stationary over relatively long periods, the spectral characteristics of the noise will be similar from frame to frame. The presence of speech could thus be detected by looking for deviations from the spectral characteristics of the background noise [Freeman et al. 89].

The weak high frequency unvoiced sounds can also be taken into consideration by tracing the zero crossing rate of the noisy signal. Besides the energy computation, an additional simple measurement is performed, the so-called zero (level) crossing rate defined as the number of times succesive samples have different algebraic signs [Rabiner & Schafer 78]. The rate at which zero crossings occur is a simple estimate of spectral properties obtained in the time domain, based on the short-time average zero crossing rate.
For a signal $x[n]$ the zero crossing rate $Z[n]$ is determined as follows:

$$Z[n] = \sum_{m=-\infty}^{+\infty} |\, sign(x[m]) - sign(x[m-1]) \,| \; w[n-m] \tag{7.1}$$

where

$$sign(x[m]) = \begin{cases} 1 & \text{if } \; x[m] \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{7.2}$$

and

$$w[m] = \begin{cases} \dfrac{1}{2N} & \text{if } \; 0 \leq m \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

Usually N is chosen to correspond to a window of 10 ms duration. Similarly to the energy computation, the zero crossing rate can also be computed at a reduced sampling rate, e.g. every 10 ms.
The energy of voiced speech is concentrated below 3 kHz, whereas the unvoiced speech energy is concentrated at high frequencies. For unvoiced speech, the mean short time average zero crossing rate per 10 ms is about 30, for voiced speech it is about 5 [Rowden 91]. As unvoiced speech has generally low energy, the zero crossing rate is a good measure for detecting unvoiced speech [Rabiner & Sambur 75] e.g. in noise. From a table containing a range of values for the energy and zero crossing rate for voiced speech, unvoiced speech and silence, presented in [Al-Hashemy & Taha 88], it can be concluded that in low energy regions the zero crossing rate is the decisive measurement for detecting speech.

The linear prediction residual or linear prediction error is another common feature used in speech detection, furthermore, it can also be used in the classification of speech as voiced or unvoiced. Feeding into the VAD the linear prediction residual instead of the input speech signal it was found that the accuracy of the VAD decision had been improved in almost all cases [El-Maleh & Kabal 97].

Different VAD algorithms have been standardized such as the one described in [ITU Rec. G.729]. Here a set of difference parameters concerning the full band energy, the low band energy, the zero-crossing rate and a spectral measure (line spectral frequencies derived from the linear prediction coefficients) is extracted. These parameters will be used for a multi-boundary initial decision in the space of the four difference measures. If none of the fourteen boundary decisions is true, the initial voice activity decision is set to 0. The final decision is obtained after smoothing, thus reflecting the long-term stationary nature of the speech signal. The set of differential parameters is obtained at each frame and represents a difference measure between the current frame parameters and the running averages of the background noise characteristics. These averages are updated only in the presence of background noise using a first order *Auto-Regressive (AR)* scheme. Different AR coefficients are used for different parameters and different sets of coefficients are considered when a large change of the background noise characteristics had been detected. Initialization of the running averages of the background noise is performed during the first 320 ms for which it is assumed that only noise is present.

For a mobile environment, the biggest difficulty lies in detecting low level speech in the presence of a range of different types of high background noise. When parts of the speech utterance are buried below the background noise, it is very hard to distinguish between speech and noise using only simple level detection algorithms. Under these conditions, spectral characteristics of the input signal must also be taken into consideration.

The GSM VAD [GSM Rec. 06.32 95] incorporates an inverse filter and an adaptive threshold which are updated during noise-only periods. As it is dangerous for the VAD to update the inverse filter autocorrelated predictor coefficients and the threshold on the basis of its own decision, a secondary VAD is used to provide the speech/noise decision for the update periods. The secondary VAD makes its decision based on the spectral distorsion between consecutive frames. If this distorsion is below a defined threshold for a sufficiently long period of time and if no pitch component is detected, it is assumed that no speech has been detected and the coefficients and the noise dependent threshold can be updated. The simplified block diagram of the algorithm is presented in Figure 7.1. If speech is present, the noise is attenuated by the inverse filter, leaving mostly deviations from the spectral characteristics of noise which are assumed to be speech. The energy of the filtered signal is compared to the noise dependent threshold. If the energy is greater than the threshold, speech has been detected. To eliminate

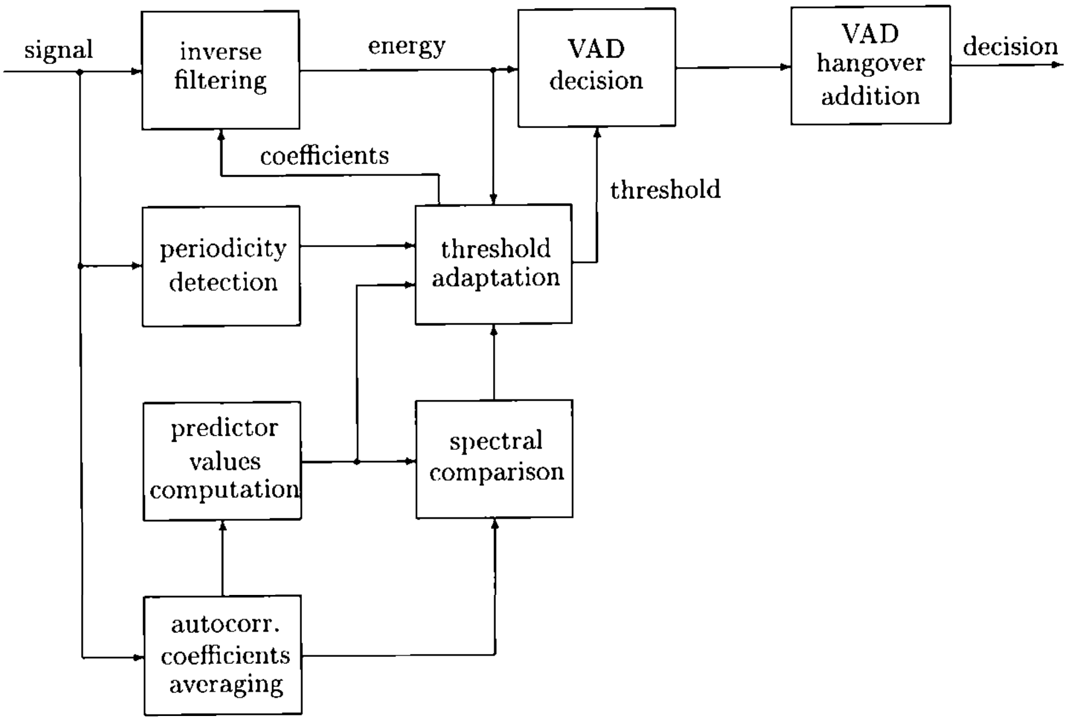


Figure 7.1: GSM voice activity detector block diagram

mid-burst clipping of low level speech, a hangover period of five frames is added, presuming the speech burst is at least three frames long.

The major weakness of the GSM VAD is found to be its assumption of the stationarity of background noise [El-Maleh & Kabal 97], which is not always true in mobile environments. Also, its performance deteriorates for SNR below 20 dB. Modifications to the algorithm improving the results of the VAD are presented in [Srinivasan & Gersho 93]. These modifications are concerning the lower frequencies, where most of the mobile noise energy is present and the GSM VAD is not very effective. Therefore a multiband energy comparison is proposed, with the goal of increasing the sensitivity of the VAD at low frequencies. This scheme compares the energy levels in four different subbands to corresponding adaptive thresholds. If any one of these thresholds is exceeded, the presence of speech will be indicated.

7.2 Double-Talk Detection

Double-talk is defined to be the period of time when both the far-end and the near-end speakers are active. During this situation the signal captured by the microphone consists of the near-end talk and the disturbing echo of the far-end speaker. In the acoustic echo cancellation topic, double-talk situations have to be detected in order to prevent misadjustment of the echo canceller. If the adaptation process is not disabled when double-talk occurs, the filter diverges and the result are audible clicks and pops in the output speech.

The simplest double-talk detector compares the signal level of the near-end speech to the reference far-end speech [Ganapathiraju & Picone 97]. If the far-end signal level is much higher than the microphone signal, then there is no near-end activity and the acoustic echo canceller is allowed to filter and adapt as well. After detecting near-end speech, adaptation will be stopped.

The short-time estimate of the error signal $e[n]$ can be also used by itself for detecting double-talk [Kuo & Pan 93], [Kuo & Pan 94]. After finding the maximum P_{max} and minimum P_{min} of the error function during a certain period of L samples, double-talk is detected when the error level satisfies the following conditions:

$$P_{error} > (0.5P_{mm} \; + \; P_{min}) \tag{7.4}$$

which means that double-talk has started and

$$P_{error} \leq (0.5P_{mm} \; - \; P_{min}) \tag{7.5}$$

when double-talk ends. P_{mm} is defined as the difference between P_{max} and P_{min}.

Another simple measure for detecting double-talk is dealing with the ratio [Johnson et al. 90] defined in Eq. (7.6). If the condition

$$\frac{|\overline{y[n]\;x[n]}|}{\overline{x^2[n]}} \geq \varepsilon \tag{7.6}$$

is satisfied, it can be considered that double-talk is present. ε is a small positive constant related to the allowable parameter estimation error.

The correlation measure between the loudspeaker and the microphone signals is a good indication of the occurrence of double-talk. In [Heitkämper 94] and [Heitkämper 97] the short time estimates of the far-end signal $\bar{x}_s[n]$ and of the error signal $\bar{e}_s[n]$ are computed using first order recursive filters with different time constants depending on whether the corresponding signal is rising or falling. A small time constant for the rising signal allows a fast tracking of the beginning of speech sequences [Heitkämper & Walker 93].
The double-talk situation is characterized by an increase of the error signal above the estimated echo caused by the far-end speaker. Thus, the crosscorrelation $\rho_{xy}[n,l]$ between the loudspeaker and the microphone signal is an indication whether the microphone signal is mainly caused by the far-end signal:

$$\rho_{xy}[n,l] = \frac{|\sum_{i=0}^{N} x[n-i]\;y[n+l-i]|}{\sum_{i=0}^{N}|x[n-i]\;y[n+l-i]|} \tag{7.7}$$

where N is chosen equivalently for 50 ms, l will be varied in the range of about 10 ms, in order to include the expected delay of the direct sound wave of the system, i.e. the main delay of the echo signal due to the echo path. n represents the time instant. If this measure is tending to 1, only the far-end speaker is considered to be active and a coupling factor

$$c_{lm}[n] = \frac{\bar{e}_s[n]}{\bar{x}_s[n]} \tag{7.8}$$

at time instant n can be calculated and updated, respectively . The product $c_{lm}[n]\,\bar{x}_s[n]$ is an estimate of the error signal when no near-end activity is present. The coupling factor will be updated only in the absence of double-talk, i.e. when the estimated error signal originating from the loudspeaker is equal to or greater than the short-term average magnitude of the error signal.
Thus the condition for detecting double-talk is the presence of far-end activity and a stronger error signal than the estimated echo originating from the loudspeaker signal [Heitkämper 97].

Spectral measures such as the Itakura-Saito or the cepstral distance can also be used as double-talk detectors [Boudy et al. 95]. These measures compare the microphone signal composed of the near-end speech, the acoustic echo and the background noise, to the loudspeaker signal. A sudden increase in the measure implies the presence of double-talk or background noise.

The double-talk situation can be detected in the frequency domain as well. The *magnitude squared coherence (MSC)* computed between the microphone and the loudspeaker signal is considered to be a useful measure in detecting double-talk situations [Le Bouquin-Jeannès et al. 96], [Gänsler et al. 96]. Because of the difficulty of identifying the echo in the presence of noise, a noise reduction will be first performed on the microphone signal, and the coherence function will then be calculated between the filtered microphone signal and the loudspeaker signal. On each block k the MSC will be averaged over all the frequencies. It was experimentally found that a separate averaging of the numerator and the denominator of the MSC leads to a more significant measure:

$$\frac{\sum_{f\in F} |\gamma_{y_{filt}x}[f,k]|^2}{\sum_{f\in F} \gamma_{y_{filt}y_{filt}}[f,k]\ \gamma_{xx}[f,k]} \tag{7.9}$$

where $\gamma_{y_{filt}x}[f,k]$ is the cross *power spectral density (psd)* between the filtered microphone signal y_{filt} and the loudspeaker signal x, $\gamma_{y_{filt}y_{filt}}[f,k]$ and $\gamma_{xx}[f,k]$ are the psd's of y_{filt} and x. F represents the set of frequencies and k is the block index. If the noise is sufficiently reduced, this measure will be close to 1 for exclusive far-end activity and will decrease in any other situation.

The double-talk problem can also be considered from a different point of view. The detector presented in [Ye & Wu 91] does not actually detect the double-talk periods, but rather decides whether the adaptive filter has converged or not. The detector is based on the principle of orthogonality, which states that after convergence to the optimal solution the following equation is fulfilled:

$$E[e[n]\,\mathbf{x}[n]] = \mathbf{0} \tag{7.10}$$

After convergence, the adaptation is halted in order to protect the adaptive filter from being disturbed by double-talk interference. If the adaptive filter has not converged yet or the echo path has changed, the adaptation will continue. To distinguish the echo path variations from double-talk situations, the average crosscorrelation between the loudspeaker signal $x[n-i]$ and the error signal $e[n]$ is calculated. The crosscorrelation coefficients are updated using an exponentially weighting recursive algorithm with an weighting factor $0.9 < \lambda \leq 1$. Whenever this average value exceeds a certain properly chosen threshold, the detector decides that the acoustic echo canceller had not converged yet or that a change in the echo path had happened. The adaptation process will then continue. Otherwise, if the average crosscorrelation is less than the chosen threshold, the adaptation is halted, thus avoiding the possibility of disturbance in a double-talk situation. The detection threshold should be chosen just a little bit greater than the average crosscorrelation value in the steady state. A too big value will make the tracking of the echo path difficult. On the other hand, a too small value will increase the misadjustment during double-talk.

7.3 Speech Detector with Variable Threshold

The combined sytem of acoustic echo cancellation and noise reduction presented in Chapter 5 is also provided with two VADs, one on the far-end, the other on the near-end side. The far-end VAD will control the acoustic echo canceller, while the near-end VAD's noise estimate output will be used in the stepsize control algorithm presented on page 96 and in the noise reduction algorithm described in Chapter 8.
In accordance to the suggested simplicity and ease of implementation of the proposed system, an energy based voice activity detection will be performed.

The new voice detection algorithm developed and tested in this dissertation makes use of an energy based adaptive threshold.
After an initial period of 100 ms, when it is assumed that the speaker has not yet started to talk, an initial background noise estimate is computed. This estimate will be the starting-point for the definition of the adaptive threshold. The just

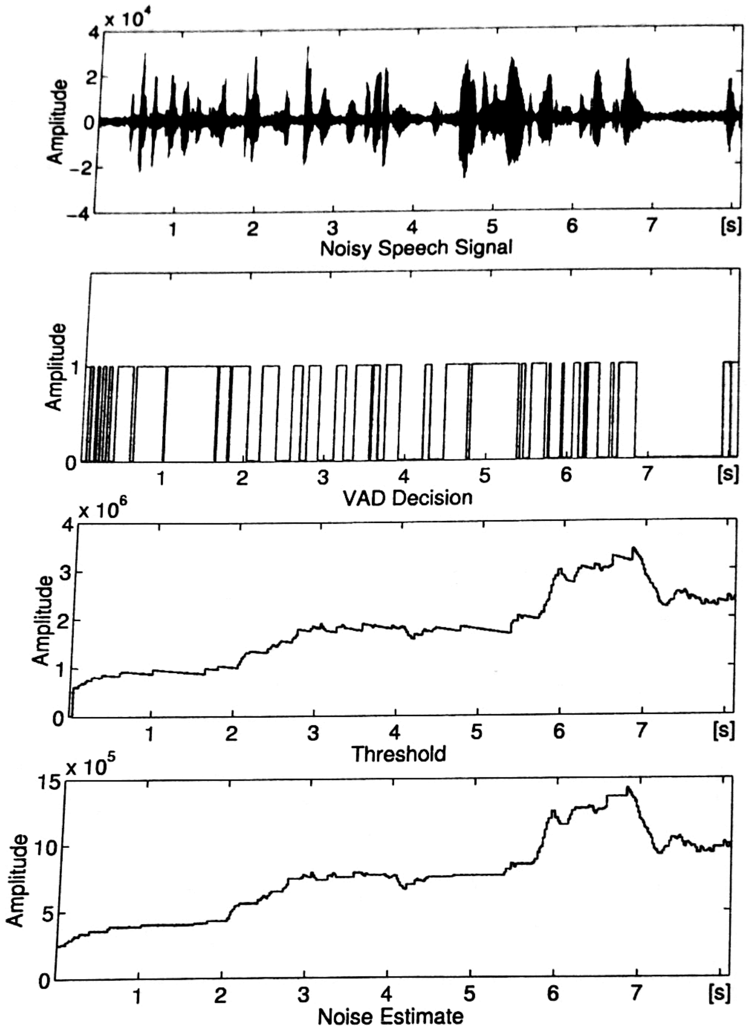
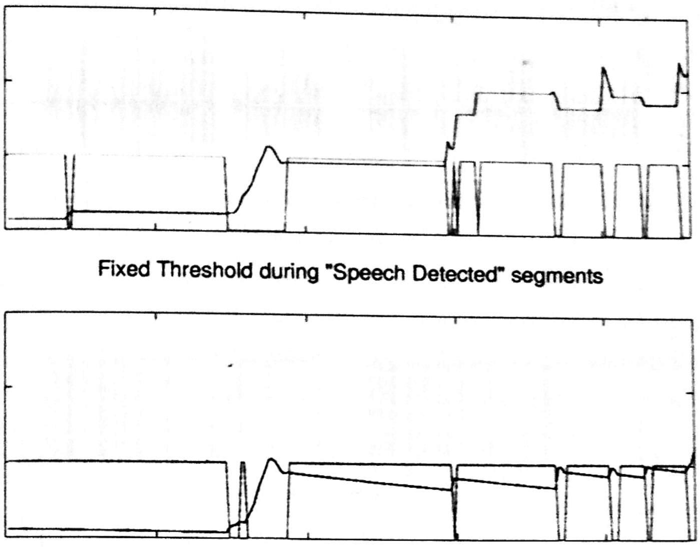


Figure 7.2: VAD results for a noisy speech signal of SNR = 0 dB

estimated background noise will be multiplied by a factor k_{on}, the VAD thus being prepared to detect a speech event.

Depending on the current decision of the VAD, two different multipliers will be used for defining the adaptive threshold. Accordingly, during speech pauses the estimated background noise will be multiplied by the factor k_{on}. This threshold has to be exceeded by the signal energy in order to detect speech.
When speech activity is detected, the adaptive threshold will be gradually decreased by applying a factor $k_{off} < 1$, but very close to 1, after every new "speech detected" decision of the VAD. This procedure leads to a slowly decreasing threshold during periods of speech, thus prolonging this period and making an hangover unnecessary.



Fixed Threshold during "Speech Detected" segments

Decreasing Threshold during "Speech Detected" segments

Figure 7.3: VAD results with fixed and decreasing threshold during speech segments

Figure 7.3 shows the difference in the VAD decision for both cases of constant and continuously decreasing thresholds during the periods when speech had been detected. It can be observed that short gaps in the decision with fixed threshold can be overcome by slowly decreasing the adaptive threshold. The proposed threshold updating rule will imply an extension of the detected speech period and therefore it will not be necessary to introduce extra hangover periods for taking into account the possible low-energy unvoiced speech at the end of the utterance.

During speech-free periods, the noise estimate will be updated by applying a first order recursive filter. In this way the long-term stationarity of noise will be taken into account.
The threshold, which in fact is the product between a multiplier (k_{on} or k_{off}) and the estimated background noise energy will be updated continuously, once every frame. This update procedure is necessary to be performed continuously, because during speech pauses, the estimated noise level will vary on frame basis. The adaptive threshold, as a function of this noise estimate will follow the background noise estimate. During detected speech periods the adaptive threshold will also be updated on a frame basis. To be precise, it will be gradually decreased, the reason being the necessity of not missing the low energy ending of speech utterances. These updating actions can be traced on the threshold variation curves presented in Figures 7.2 and 7.4, which show the VAD relevant signals for two noisy speech signals with different SNRs.

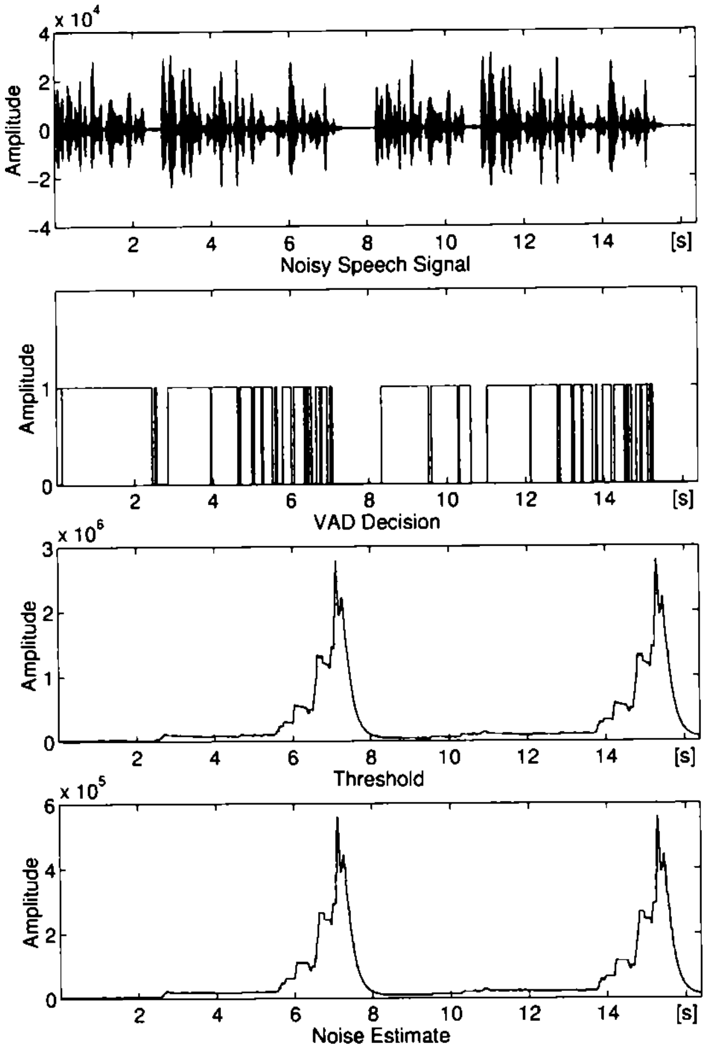


Figure 7.4: VAD results for a repeated sequence of speech signal (SNR = 20dB)

The essence of the above described algorithm can be presented by the following C code lines:

```
for (i=1; i<MaxFrameCount; i++)
{
  if (VAD_Decision == 0)
  {
     NoiseEstimate[i] = NoiseEstimate[i-1] * Beta +
                        (1-Beta) * CurrentFrameEnergy;
     AdaptThresh[i]   = NoiseEstimate[i] * k_on;
  }
  else
     AdaptThresh[i]  = AdaptThresh[i-1] * k_off;
}
```

where the variable NoiseEstimate[i] represents the estimated background noise energy for frame i, Beta is the lowpass filter coefficient, typically set to 0.9 and CurrentFrameEnergy represents the energy of the current frame.

A very important issue in the voice activity detection topic is the possibility of accumulating errors during noise detection. In such cases, what seem to be very long speech pauses or continuous speech activity are very likely the result from failures in the detection. This happens when a too high or a too low threshold cannot be correctly updated so as to permit a proper function of the detection algorithm. Therefore the algorithm proposed in this work was tested on a signal of double length consisting of two identical speech segments. The same decisions for both segments had to be expected. The results presented in Figure 7.4 confirm the expectations.

The double-talk detector implemented in this work is based on the algorithm presented in [Johnson et al. 90] and described by Eq. (7.6). Figure 7.5 presents the loudspeaker signal, the microphone signal (composed of the echo signal and a near-end speech signal) and the output of the double-talk detector. It can be observed that the DTD decision is set to one only during periods when near-end speech is present at the microphone input. Thus double-talk segments can be detected. During near-end periods of silence the DTD yields zero at its output. In this case the microphone input signal will be containing only the echo signal resulting from the emitted loudspeaker signal.

As already mentioned in Chapter 5, when speech is present on the far-end side, this information being supplied by the far-end VAD, the actions performed by the AEC will be determined by the output of the DTD:

- if no double-talk is detected, the AEC performs normally, filtering and adaptation of the transversal filter coefficients takes place
- if double-talk is detected, the filter coefficient adaptation process is stopped, while filtering still continues. Thus the filter is prevented from divergence.

The output of the DTD is irrelevant if there is no speech radiated by the loudspeaker into the vehicle interior.

7.4 Conclusions

The function of voice activity and double-talk detectors is very important in acoustic echo cancellation as well as in noise reduction systems, where there is need of an algorithm that has the ability to decide whether a signal segment contains speech or only noise. Thus, the AEC has to stop its adaptation when

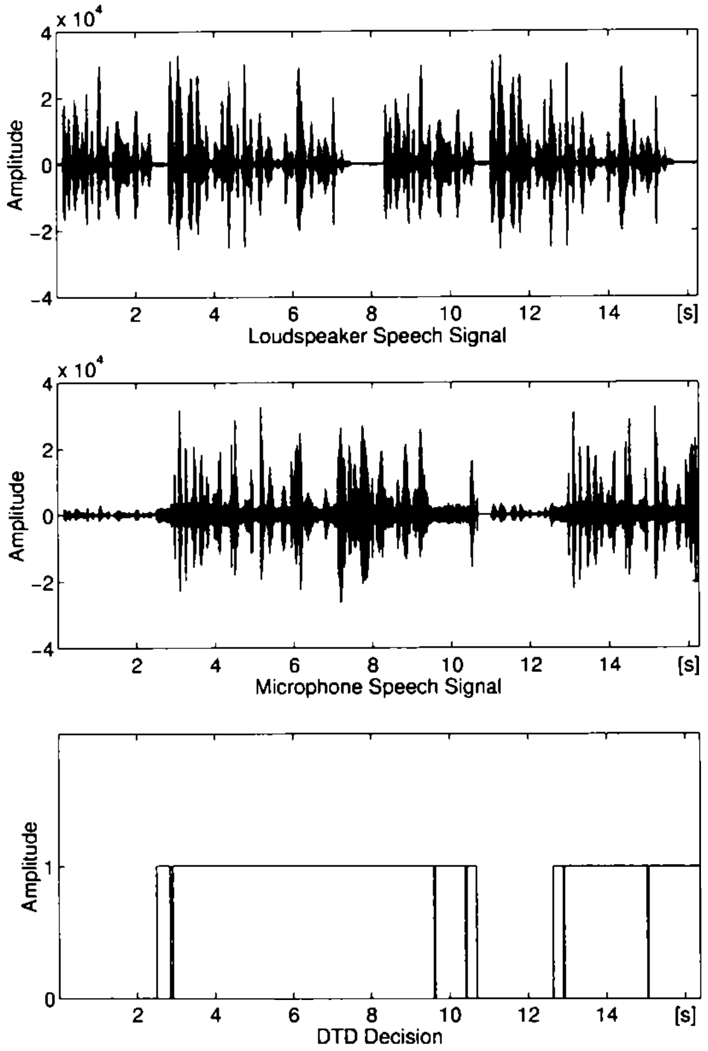


Figure 7.5: DTD results

the VAD applied to the far-end signal decides that there is no speech present. In this situation the double-talk detector must control the filtering process of the AEC. If a near-end speaker is active, the microphone signal is no longer allowed to be filtered, otherwise the near-end speech will be distorted by the AEC.

In the field of speech enhancement, the background noise characteristics can be learned during speech pauses. That is why VADs are also necessary in noise reduction systems, especially in those with only one microphone such as the new algorithm presented in Chapter 8.

After presenting the most commonly used speech detectors, a new VAD has been proposed in this chapter. It uses an adaptive threshold which will be continuously updated, during speech pauses as well as during speech activity. Only the update procedure differs.

During speech pauses the estimated background noise will be updated and at the same time the adaptive threshold which is a factor (k_{on}) times the estimated noise. When speech activity is detected, the adaptive threshold will be gradually decreased by applying a factor $k_{off} < 1$, but very close to 1, after every new "speech detected" decision of the VAD. This procedure leads to a slowly decreasing threshold thus prolonging the detected speech period. The reason for introducing this new updating rule during speech periods was to avoid the necessity of hangover periods at the end of an utterance, which are meant to prevent the possible low-energy unvoiced sounds from being considered as noise.
The new algorithm thus performs a reduction of computational complexity.

Chapter 8

Noise Reduction System

Speech enhancement methods, as presented in Chapter 3, can be classified according to the number of microphones used and the domain of implementation, which can be either the time or the frequency domain.

Single microphone systems have the advantage of using simple standard recording equipment, but the algorithm is able to cancel only stationary noise. The method makes use of a speech/noise or voice activity detector. It assumes that the noise estimate calculated during speech pauses is valid during periods of speech as well. Rapidly varying noise can thus cause problems. Systems with two or more microphones need more hardware and some knowledge about the place of the desired source, but therefore permit cancellation of nonstationary or very strong interfering noise. As handsfree equipments are consumer products, the "one microphone approach" is preferred by the manufacturers.

Speech enhancement algorithms can be implemented in the time domain or in the frequency domain. The decision for a time domain implementation had been taken in Chapter 5. Any transformation into the frequency domain is associated with a time delay needed for the processing of an FFT. The time limitation, given by the GSM requirements of no more than 39 ms of delay allowed for AEC and NR processing, leads to the challenging field of time domain noise reduction. When only one noisy observation is available, an effective and robust speech detector plays an important role in noise suppression systems.

8.1 Noise Reduction Algorithm in the Time Domain

In this thesis the "one microphone approach" and the time domain implementation are considered. The noise reduction system included in the proposed combined system is the first projection order of the APA, in the direct linear transversal form with a filter length of 256. The implemented structure is shown in Figure 8.1.

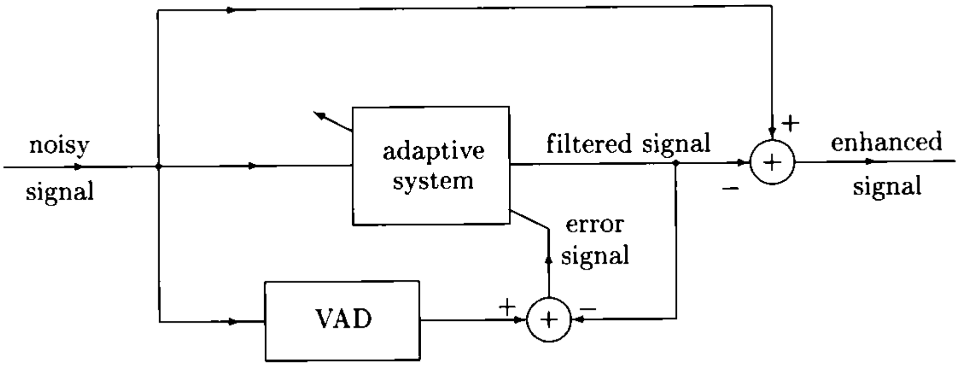
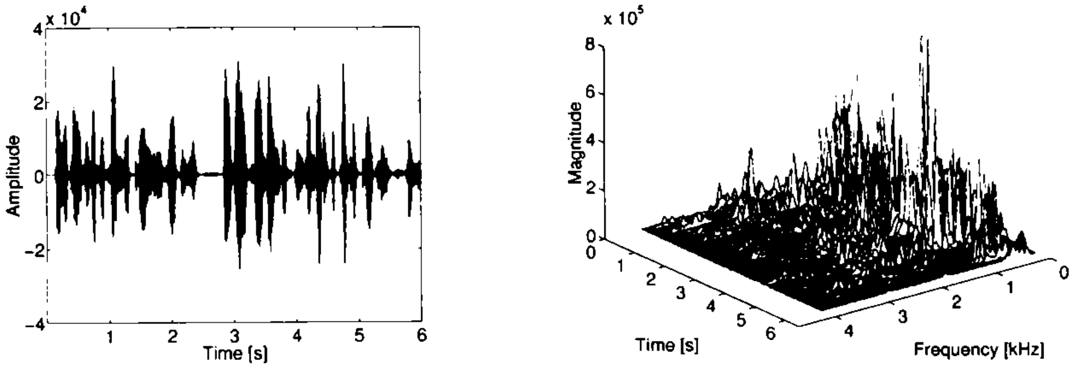


Figure 8.1: The proposed noise reduction system

The estimated noise, delivered by a voice activity detector, is applied to the desired signal input of the algorithm, while the noisy signal is considered to be the reference input signal of the NLMS. The algorithm will adapt so as to output a filtered signal that is as close as possible to the desired noise estimate. The filtered signal will then be subtracted from the noisy input signal, thus supplying an enhanced signal at the noise reduction system's output.

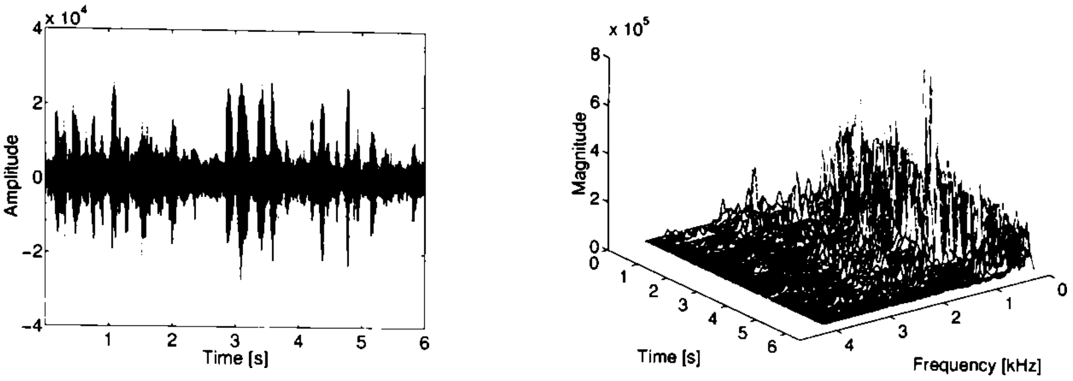


a. time domain

b. spectrogram

Figure 8.2: Clean speech signal

As already mentioned, the quality of the estimated noise signal is decisive. The noise reduction stands or falls by the VAD estimated background noise. During experiments the noise signal was known (See section 5.4.), as it was the signal added to the clean speech in order to get the noisy speech signal. Considering the clean speech signal as represented in Figure 8.2 and the noisy signal from Figure 8.3, two possible results given by the proposed noise reduction system are

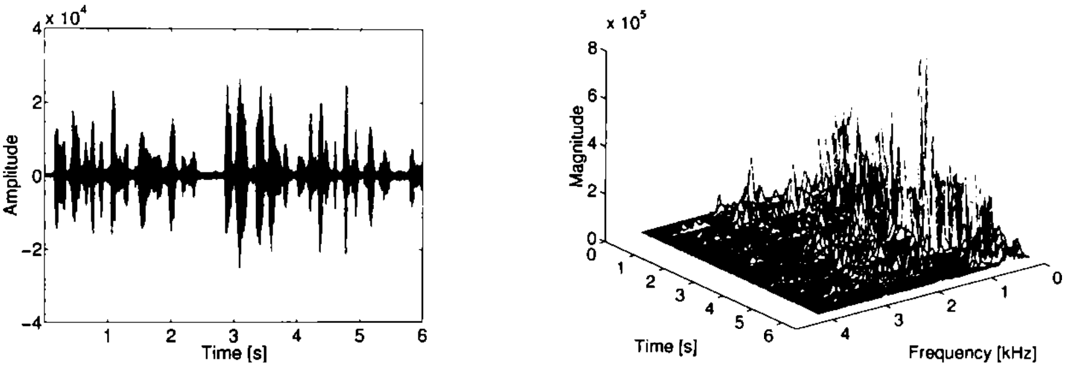


a. time domain

b. spectrogram

Figure 8.3: Noisy speech signal (SNR = −10 dB)

shown in Figures 8.4 and 8.5.



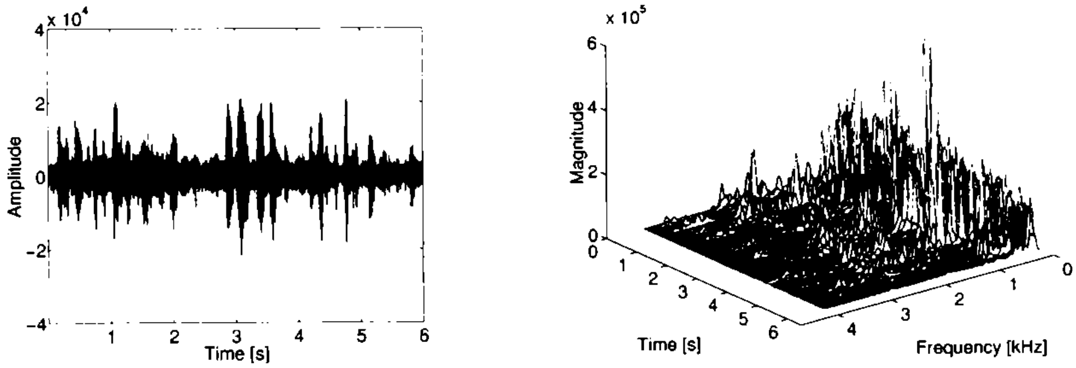
a. time domain

b. spectrogram

Figure 8.4: Noisy speech signal of Figure 8.3 enhanced by the proposed system with known noise reference

Figure 8.4 represents the best case of speech enhancement achievable by the proposed noise reduction system, namely when using an ideal VAD for the background noise estimation. The ideal VAD can be approximated by applying the known noise signal to the algorithm instead of the estimated noise signal. It can be seen that the noise could be almost entirely removed without distorting the speech signal. The enhanced waveform is very close to that of the clean speech signal (Figure 8.2). A minimal remaining amount of noise can be observed only during the short speech-free segments.

The speech enhancement performance when using a simple energy based VAD, as the one presented in section 7.3, is represented in Figure 8.5. An overall attenuation of the noisy signal can be noticed. The noise reduction is minimal and



a. time domain

b. spectrogram

Figure 8.5: Noisy speech signal of Figure 8.3 enhanced by the proposed system with estimated noise reference

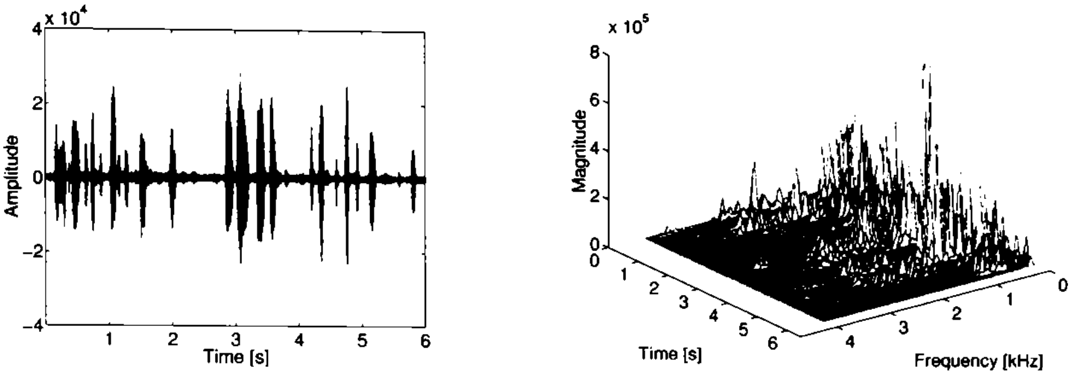
obviously not acceptable for a noise reduction system in a noisy mobile environment.

The Itakura-Saito distorsion measure, described in section 5.3.2 as an objective measure characterizing the similarity of two signals, can give more information about the quality of the enhancement procedure. The noisy speech signal of Figure 8.3 shows a mean Itakura-Saito distance to the clean signal of 2.2. The best possible enhancement made available by the proposed noise reduction system will yield an Itakura-Saito distance of 0.9.
When using the simple energy based VAD, the performance of the NR system will lead to a distorsion measure of 2.0. This is not a satisfying result for the proposed system. The reason lies in the insufficient noise estimation performance of the VAD algorithm. More complex detectors would certainly deliver better results.

8.2 Comparison to Noise Reduction in the Frequency Domain

The frequency domain has also been considered. The linear and nonlinear spectral subtraction as well as the spectral scaling have been implemented and tested. In the following the time and frequency domain results due to nonlinear spectral subtraction are presented and compared to the proposed time domain noise reduction system based on the APA of dimension one.

Analyzing Figure 8.6, which represents the enhanced signal after nonlinear spectral subtraction and comparing it to Figure 8.2 it can be observed that in the time domain representation the waveform is partially completely changed.



a. time domain

b. time and frequency domain

Figure 8.6: Noisy speech signal of Figure 8.3 enhanced by Nonlinear Spectral Subtraction

The background noise had been removed satisfactorily but to the detriment of speech distorsion. As to the frequency domain graph, not only the low frequency but also higher frequency components belonging to the speech signal have been partially removed.

When listening to the enhanced signal it can be noticed that the background noise is almost totally removed, but there are some annoying musical tones present which cannot be ignored. The Itakura-Saito distance between the enhanced signal and the clean speech signal supplies a value of 3.2. This means a rather distorted signal after noise reduction.

It is possible to find a compromise between the amount of noise being removed and the musical tones. Setting the noise floor to a higher value will leave more noise in the enhanced signal but the musical noise will be less disturbing.

8.3 Conclusions

When using handsfree operation in the noisy environment of a moving car, there is need of a speech enhancement system. The background noise from sources such as road, wind, fan or tyres leads to a reduction in conversational speech quality. Noise reduction systems intend to achieve an increase in intelligibility keeping the distorsions introduced by the enhancement algorithm at an acceptable, not annoying level. The naturalness of the residual noise is very important. Additive noise can be suppressed by capturing it separately from the desired speech and subtracting it from the noisy signal. This requires a second microphone for providing the noise reference.

Because of the low cost restriction of handsfree equipment, only one microphone will be available, analysis during speech pauses will furnish the required noise estimate. Single microphone speech enhancement systems make use of a speech/noise

or voice activity detector for estimating the background noise. This method leads to less enhancement than the two microphone system, since it assumes that noise during speech pauses is representative of noise during periods of speech. Thus, rapidly varying noise can cause problems.
As mentioned in Chapter 3, due to the psychoacoustic properties of the human ear, modifications to the noisy signal are best performed in the spectral domain. As any transform into the frequency domain is connected to a time delay and because of the GSM restriction concerning digital processing delays, in this thesis the time domain implementation will be investigated.

The proposed noise reduction system works in the time domain on a sample-by-sample basis. The algorithm is based on the first order affine projection algorithm, i.e. the NLMS algorithm. The estimated noise, delivered by a voice activity detector, is applied to the desired signal input of the algorithm, while the noisy signal is considered to be the reference input signal of the NLMS. The algorithm will adapt so as to output a filtered signal that is as close as possible to the desired noise estimate. This signal will then be subtracted from the noisy input signal.
The algorithm needs a very good estimate of the background noise in order to achieve acceptable results. A standard low complexity VAD is not sufficient, it must be a powerful algorithm. The experimental results with the known background noise lead to very good noise reduction performance, but when implementing the simple energy based VAD the enhancement is minimal and unacceptable for a mobile environment.

This leads to the conclusion that either a very sophisticated VAD must be available or the noise reduction has to be performed in the frequency domain where the human ear is not so sensitive to distorsions introduced by a nonideal noise estimate.

Chapter 9

Summary

The main goal of this thesis is the investigation of algorithms concerning acoustic echo compensation and noise reduction and the realization of a combined system suitable for car handsfree applications in the GSM network. The system is completed by the use of voice activity and double-talk detection algorithms which are very important for the correct operation of echo cancellation and noise reduction. As the combined AEC and NR system will be running on a digital signal processor, the computational complexity of the proposed algorithms has to be reduced and the algorithms must be suited for fixed-point implementation.

For a comfortable handsfree communication, the acoustic echo generated by the loudspeaker-room-microphone system has to be suppressed. The acoustic echo canceller provides an adaptive estimate of the room impulse, the echo compensation is performed by subtracting this synthetically generated estimate from the microphone input signal. In a car environment a second impairment exists, namely the omnipresent background noise from sources such as road, wind, fan or tyres. The perceived effect of this additive noise is a reduction in speech quality. For a better speech intelligibility noise reduction is very important.

The exact knowledge of the environmental conditions in which GSM car handsfree systems operate is crucial in the development of combined acoustic echo cancellation and noise reduction systems. As the acoustic echo canceller has the task of suppressing the echo generated by the loudspeaker-room-microphone system, first a digital replica of the LRMS must be defined (*Chapter 4*). Considering the peculiarities of the time-varying nature of the car cabin environment, the applicable algorithms have to be adaptive, which means that the adaptive filter has to be a selfdesigning device capable of tracking the echo path variations. The tranversal FIR filter in the linear direct form has been proposed to be used because of its robustness and ease of implementation. IIR filters, although computationally less demanding, are much more difficult to handle, especially when the filter inherent feedback has to be combined with the adaptation process. Concerning the adaptation algorithm, it has been suggested to perform a stochastic gradient algorithm which is numerically robust and computationally less demanding than the least-squares algorithms.

The two separately working systems of AEC and NR can be merged into a single symbiotic system. A new combined system is proposed in *Chapter 5* having its elements working entirely in the time domain. This has been considered because of the time constraint of 39 ms of processing time for both acoustic echo compensation and noise reduction imposed by the GSM specifications. A noise reduction system in the frequency domain, as it is usually considered, would need a Fourier Transform implementation which, when the transform is performed over 256 samples already requires 32 ms of processing time and also more computational power than a time domain implementation. Another disadvantage of noise reduction performed in the frequency domain are the so-called musical tones that appear because of the assumption of short-time spectral stationarity of the noisy speech signals and the nonstationarity of the noise. As no reevaluation during speech periods is done for the noise estimate computed during speech pauses, this estimate will be no longer valid if the noise varies rapidly. A noise reduction algorithm in the time domain operating on a sample-by-sample basis would be computationally less demanding and would save processing time.

After implementation and testing of many of the existing algorithms, it was found that the affine projection algorithm (APA) in its fast realization is well suited for the acoustic echo compensation task. For speech input signals it converges faster than the well-known NLMS algorithm. The additional computation effort is reasonable. In *Chapter 6* the adaptation procedure in the proposed combined system is discussed. An AEC using the APA is proposed which additionally has been made more robust to near-end noise by designing a new stepsize control depending on the ratio between the loudspeaker signal power and the estimated background power. In high background noise the adaptation coefficient will be close to zero thus slowing down the adaptation process. The simulation results confirm this assumption. An improvement of the Echo Return Loss Enhancement (ERLE) was achieved for every of the tested projection orders of the APA.

The performance of the APA has been investigated in the fullband and subbands approach as well. The filter bank implementation seemed to be worth examining because of the complexity reduction it promises. As a conclusion to the results obtained in noiseless and noisy near-end environments in the subband approach, it can be suggested to choose an affine projection algorithm of dimension 4, where the higher the filter order the better the ERLE and convergence performance. Higher dimensions of the affine projection algorithm do not justify the increase in computational complexity.
For the fullband approach, which has been considered in this thesis, the affine

projection algorithm of order 4 combined with the new adaptive stepsize control can be recommended for both noiseless and noisy near-end environments. Higher orders of the adaptation algorithm do not justify the extra computational requirements and do not offer better results.

Another important issue in the handsfree topic is the reliable detection of the presence of speech. Therefore in *Chapter 7* a reinforced attention had been paid to voice activity detectors (VAD) which have the task of preserving the echo compensation algorithm from adapting on a false signal and thereby distorting the signal picked up by the microphone. Another important application of speech detectors is their importance in speech enhancement because of their ability to estimate the background noise during speech pauses. In accordance to the suggested simplicity and ease of implementation of the proposed system a new energy based variable threshold voice activity detector has been developed and tested. The update procedure of the adaptive threshold differs according to the current detected state of "speech" or "noise". This algorithm represents an improvement to the existing energy based algorithms in that is needs no extra hangover processing.

As the combined system is supposed to perform entirely in the time domain, in *Chapter 8* the noise reduction system has been designed on the basis of the first projection order of the APA. Testing it with known noise estimates showed a very good performance of noise reduction. But when embedded in the whole system and getting its noise reference from a simple, energy based VAD, the quality was found to be inadequate. Therefore it was concluded, that either a sophisticated VAD has to be used or the noise reduction must be performed in the frequency domain where the human ear is not so sensitive to distorsions introduced by a nonideal noise estimate.

The following items are claimed to be original contributions of this thesis:

- The approach of considering AEC and NR entirely processing in the time domain which points to a reduction of processing delay inherent to any frequency domain implementation (*Chapter 5*)
- The treatment of the combined system as a whole including the design of voice activity detection in the context of GSM, which according to the studied literature until now has been considered as different topics (*Chapter 5*)
- The theoretical derivation of the noise-dependency of the affine projection algorithm, which has been confirmed by investigations made in real noisy near-end environment (Eq. (6.24))
- The investigations made on the APA in noisy environments, which in this form have not yet been presented in the literature (*Chapter 6*)

- The subband investigations made on the APA in noisy environment: for the filter bank approach there was proposed to use a subband implementation of a maximum APA dimension of 4 (section 6.3.2)

- The new stepsize control algorithm proposed for the enhancement of the APA performance in noisy environments, which permits a good improvement of the ERLE performance (Eq. (6.25) in section 6.4)

- The new VAD algorithm with different variable thresholds depending on whether noise or speech had been detected, making the hangover procedure unnecessary (*Chapter 7*)

- The conclusion that it is not possible to implement a noise reduction system with one microphone in the time domain using simple VAD algorithms. A very good noise estimate for the noise reduction system is a must in this case (*Chapter 8*)

- In the decision process for an adaptive algorithm and a noise reduction system with the appropriate speech detectors, a vast bibliographic investigation has been made. This was finalized by setting up a data base containing the main information to the studied topics

- For testing the algorithms under research, a library of special simulation blocks was made available, completing the existing general blocks of the test environment and permitting the future realistic investigation of the algorithms of interest

Appendix A

Basics of Speech Signals

The speech waveform is an acoustic sound pressure wave that originates in the human speech production system. The main components of the speech system are the *lungs*, the *trachea*, the *larynx*, the *pharyngeal cavity*, the *oral cavity* and the *nasal cavity*. The pharyngeal and oral cavities are usually grouped and referred to as one unit, the so-called *vocal tract*.
Two main methods by which speech sounds are produced [Waters 91] can be distinguished:

- by voicing, when the vocal cords located in the larynx are vibrating at a constant frequency, thus generating the vowels

- by the turbulent flow of air at some point of constriction in the vocal tract, which gives arise to unvoiced sounds like the consonants

The vocal tract shape causes certain frequencies in the excitation to be amplified and attenuates other frequencies, thus a set of resonant frequencies can be found. The locations of these resonances in the frequency domain depend upon the shape and dimensions of the vocal tract. Since these frequencies form the overall spectrum, they are called *formants*. The fundamental frequency will be referred to as F_0 [Rabiner & Schafer 78]. *Pitch* is another term that is often interchangeably used with the fundamental frequency. In principle, there are an infinite number of formants in a given sound, but in practice usually only 3-5 will be found in the Nyquist band after sampling [Deller et al. 93].

When the vocal cords vibrate, harmonics are produced at multiples of the fundametal frequency, the amplitude of the harmonics decreasing with increasing frequency. Such voiced speech has a spectrum with energy concentrated at discrete frequencies, i.e. the fundamental frequency F_0 of the vocal folds and multiples of F_0, i.e harmonics. The average fundamental frequency for men is somewhere between 50-250 Hz, for women it is in the range of 120-500 Hz. About one-third of speech is completely aperiodic (unvoiced), resulting from a random excitation that resembles white noise, caused by air rapidly passing through a narrow constriction in the vocal tract.

The spectral characteristics of speech are time-varying, since the speech production system changes rapidly over time. Therefore speech will be divided into segments that possess similar acoustic properties over short periods of time.
Due to the limitations of the organs for human speech production and the auditory system, typical human speech communication is limited to a bandwidth of 7-8 kHz.

The smallest element of speech which indicates a difference in linguistic meaning is called *phonem*. In fact the phonem really represents a class of sounds that convey the same meaning, because a phonem will have a variety of acoustic manifestations in the course of flowing speech [Deller et al. 93]. A phonem is written between slashes, e.g. /f/ in ”free”. The phonems can be classified into vowels, diphtongs, semi-vowels as /w/ or /r/, plosives like /b/, /d/, fricatives as /f/, /s/, affricates (/tsh/ and /dzh/) and nasals /m/, /ng/.
As a speaker utters a series phonemes, each of a brief duration averaging about 80 ms [Rabiner & Schafer 78], both F_0 and vocal tract shape evolve in time, yielding a dynamic speech signal. For accurate modelling, speech analysis must be restricted to brief sections of the signal, during which the production source has approximately stationary characteristics. During a *frame* or *window* of about 10 to 30 ms, the vocal tract usually retains a relatively constant shape and the corresponding short-time speech spectrum is a good measure of the state of the sound source. Much of the time, speech is almost periodic and thus has an approximate *line spectrum*, primarily consisting of energy centered around harmonics of F_0.

The effect of the vocal cords and the vocal tract is to introduce a measure of correlation and predictability on the random, noise-like air flow from the lungs [Vaseghi 96]. A model for speech production is presented in Figure 3.5. Human speech production can be modeled as a filter (due to the vocal tract) acting on an excitation waveform [Rabiner & Schafer 78]. The input signal for the digital filter is produced either by an impulse train generator offering a harmonic rich repetitive waveform or by a random noise generator. The digital filter, with the same characteristics as the vocal tract, will have its parameters varied corresponding to the modifications of the vocal tract. The filter is thus time-varying, the rate of variation being slow, with parameters updated every 5 to 25 ms. Either of the signal sources used in the speech model will produce a broadband spectrum of energy in the frequency domain. Frequency shaping [Waters 91] is provided by the filter characteristic which consists of a curve where the various resonances of the vocal tract appear as peaks.

Certain aspects of speech waveforms are more perceptually important than others. The auditory system is more sensitive to the presence of energy than to the absence of it and tends to ignore many aspects of phase. Thus, speech coding and enhancement algorithms concentrate on accurate preservation of peaks in the speech amplitude spectrum rather than on phase relationships or energy at weaker frequencies. Voiced speech with its high amplitude and concentration of energy at low frequency, is more perceptually important than unvoiced speech for preserving speech quality. Thus, most enhancement algorithms tend to concentrate on improving the periodic portions of speech [Rabiner & Schafer 78].

Appendix B

A Short Description of the GSM System

The specifications of the Pan-European public mobile communication system were released by the *Groupe Spéciale Mobile (GSM)* of the *Conférence Européenne des Administrations des Postes et des Télécommunications (CEPT)* by the end of 1988. They cover various aspects of the system in 13 sets of recommendations [Steele 92]. The system was named after this Groupe Spéciale Mobile Committee. The main governing body of GSM is the *MoU* - Memorandum of Understanding. The MoU's basic task is to establish internationally compatible GSM networks in member countries, and to provide a mechanism to allow for cooperation between operators in respect of commercial, operational and technical issues, e.g. international roaming, global marketing, harmonisation of tariff principles, definition of accounting and billing procedures, legal and regulatory matters, time scales for the procurement and deployment of systems [GSM MoU 98b].

The GSM system provides a wide range of services and facilities, both voice and data, that are compatible with those offered by the fixed *Public Service Telephone Networks (PSTN)*, *Public Data Networks (PDN)* and *Integrated Services Digital Networks (ISDN)*. The great advantage of the GSM system is its compatibility of access for any mobile subscriber in any country that operates the system. The GSM operating countries provide possibilities for automatic roaming, locating and updating of the mobile subscriber's status.
The GSM system is a digital system operating in two paired bands, one band (890-915 MHz) for the uplink transmission from the mobile to the base station and another band, spaced at 45 MHz above it (935-960 MHz), for the downlink transmission, where the base station transmits and the mobile terminal receives. The GSM frequency band is partitioned into 124 paired duplex channels with 200 kHz channel spacing in each band. The information is transmitted in 271 kbit/s bursts.
The most important general characteristics of the GSM system can be listed as

follows [Smolka 94]:

- a narrow band transmission with *Time Division Multiple Access (TDMA)*
- full digital speech and signalling transmission
- constant envelope, continuous phase *Gaussian Minimum Shift Keying (GMSK)* modulation robust against signal fading and interference
- speech transmission with a bit rate of 13 kbit/s in fullrate and 6.5 kbit/s in halfrate mode
- data transmission at a rate of 2.4 to 9.6 kbit/s
- up to 8 speech connections per carrier frequency, respectively up to 16 calls in halfrate operating mode
- *Voice Activity Detection* and *Discontinuous Transmission (VAD/DTX)* which minimizes the battery consumption, as the mobile is transmitting only during active speech periods
- *Discontinuous Reception (DRX)* which means that the mobile receiver is on only when paging blocks are expected to arrive (sleep mode)
- frequency hopping for minimizing the interference from frequency selective fading
- equalizer for compensating the multipath reception of excess path delays of up to 16 μs

The speech encoder takes its input as a 13 bit uniform PCM signal from the audio part of the mobile station where the signal is sampled at 8,000 samples/s or from the PSTN via an 8 bit/A-law to 13 bit uniform PCM conversion. The encoded speech is then delivered to the channel encoder, specified in [GSM Rec. 05.03 95]. Using an A-law compander, the speech sample bit rate is 64 kbit/s. The speech coder reduces this bit rate to an average bit rate of 13 kbit/s for the encoded bit stream. In the receive direction, the inverse operations are performed.
In [GSM Rec. 06.10 95] the mapping between 20 ms blocks of 13 bit uniform PCM data to encoded blocks of 260 bits according to the so-called *Regular Pulse Excitation - Long Term Prediction (RPE-LTP)* coding scheme and the inverse operation, from 260 bits to 160 reconstructed speech samples, are described in detail. The codec is specified down to the bit level, thus enabling the verification of the implementation by use of a set of test sequences.

After adding extra bits for error recognition and correction, the bit rate for the speech channel to transmit will be 22.8 kbit/s, which corresponds to a block of 456 bits/20 ms. Then interleaving, encryption, burst building and burst multiplexing is performed. The TDMA frame, made up of 8 timeslots of 156.25 bits, is input to the GMSK modulator at a bit-rate of approximately 271 kbits/s. In Figure B.1 the main elements of the transmission system are presented.

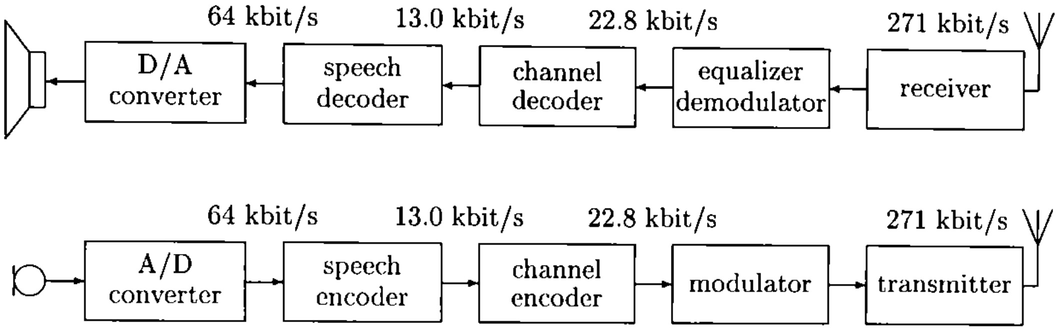


Figure B.1: The elements of the GSM communication system

The combined system of acoustic echo compensation and noise reduction proposed in this thesis will be positioned between the loudspeaker and the microphone signal path, before the D/A and after the A/D converters, the functions being performed in the digital domain.

Figure B.2 shows the positioning of the proposed combined system in the GSM system. The elements of Figure B.1 have been rearranged so that the integration of Figure 5.5 from page 64 can be easily observed. The AEC+NR system will be active only during handsfree operation mode. The microphone signal at the input of the speech encoder will be free of the acoustic echo due to the car interior reflections and the background noise will be reduced. In the handy mode, when the mobile is not placed in the cradle of the handsfree system, the reference points 1 and 2 respectively 3 and 4 will be connected directly, acoustic echo cancellation and noise reduction being bypassed. The speaker's mouth will be

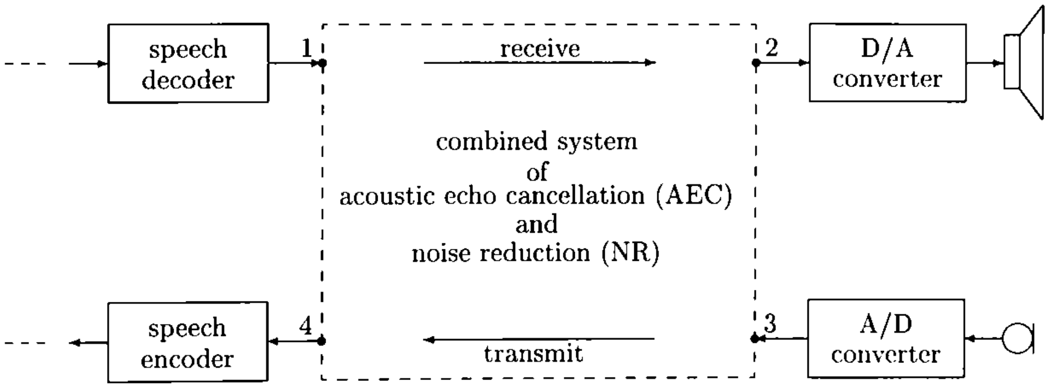

speech decoder
1
receive
2
D/A converter
combined system of acoustic echo cancellation (AEC) and noise reduction (NR)
speech encoder
4
transmit
3
A/D converter

Figure B.2: Proposed combined system in the GSM communication system

much closer to the microphone input of the mobile phone than in the handsfree arrangement, and therefore no special processing concerning acoustic echo cancellation and background noise reduction will be necessary.

The radio subsystem of the GSM system provides a certain number of logical channels which can be grouped into two categories:

- traffic channels for carrying speech and data information
- signalling channels.

These logical channels are mapped onto physical channels, defined as a timeslot, with a timeslot number from 0 to 7, in a sequence of TDMA frames. Each of the 124 paired carrier frequencies supports 8 physical channels mapped onto 8 timeslots within a TDMA frame. A given physical channel always uses the same timeslot number in every TDMA frame, i.e. one timeslot every 4.615 ms. As the GSM system also specifies frequency hopping, the physical channel can be defined as being a sequence of radio frequency channels and timeslots.

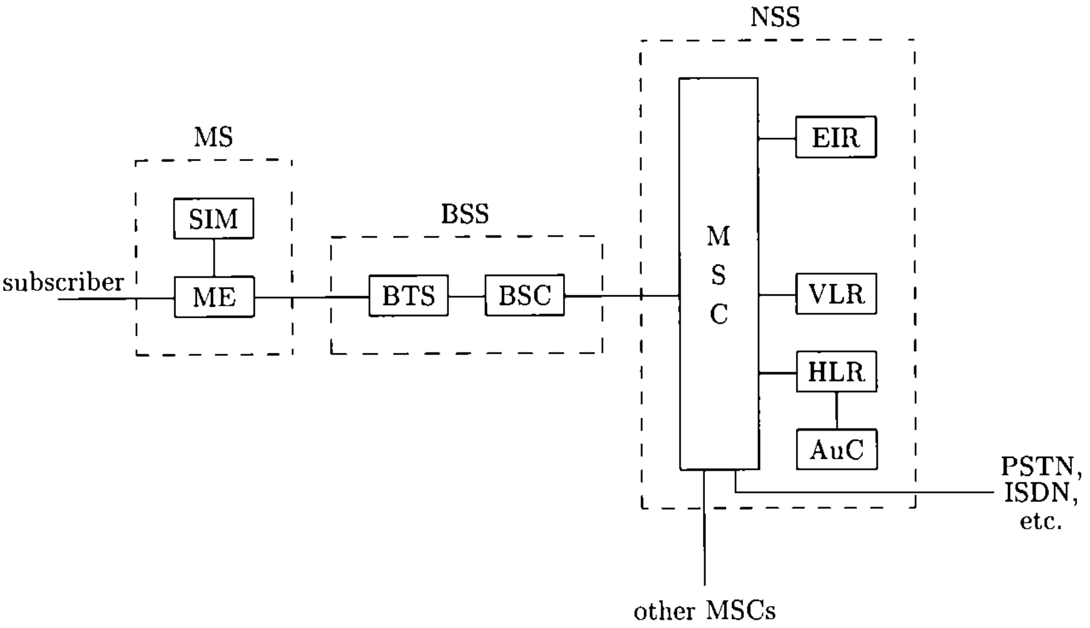


Figure B.3: GSM network architecture

Considering Figure B.3, three different subsystems can be noticed within the GSM network:

- the *Mobile Station (MS)* consisting of the *Mobile Equipment (ME)* and the *Subscriber Identification Module (SIM)* containing customer specific informations

- the *Base Station System (BSS)* consisting of a *Base Transceiver Station (BTS)* with transmit, receive and signalling units and the Base Station Controller (BSC). The BSC manages the channel assignment and the handover procedure between different cells. A BSC controls a number of BTSs. These BTSs will contact the MSs.

- the *Network and Switching System (NSS)* with the *Mobile Switching Center (MSC)*. The MSC, a main element in the general architecture of the GSM network, coordinates call setup to and from a GSM user and provides the interface with external networks [Spencer 98]. The MSC also handles mobility management, via *Home and Visitor Location Registers (HLR and VLR)* and subscriber management via the *Equipment Identity Register (EIR)* and the *Authentication Control (AuC)*.

Appendix C

A Brief Presentation of a GSM Handy

In this Annex a concise presentation of the Siemens approach to a GSM chipset is given [Siemens AG 98]. This chipset meets all performance requirements set down in the GSM recommendations for speech and data.

The Siemens HiGOLD is a complete chipset which covers all functions for a mobile terminal for GSM both for baseband and radio frequency. It is a continued development of the *GOLD (GSM One-chip Logic Device)* chipset. *HiGOLD* is the integration of microcontroller and digital signal processor in a single package which leads to a reduction of system cost, board space requirements and power consumption. The Siemens chipset is optimized for applications in very small GSM/PCN handhelds for fullrate, enhanced fullrate and halfrate services.

The fullrate chipset comprises the following chips:

- the HiGOLD chip consisting of a microcontroller part (HiGOLD-μC) and a signal processing part (HiGOLD-SP)
- the *GSM Analog Interfacing Module (GAIM)* which performs the voiceband and baseband A/D and D/A conversions and the Power Amplifier Control D/A conversion
- the RF Quadrature Demodulator Circuit (RF Receiver) for GSM/PCN/PCS1900
- the RF Quadrature Modulator Circuit (RF Transmitter) for GSM/PCN/PCS1900
- the RF PLL Circuit

For advanced firmware features such as halfrate codec or enhanced fullrate codec a coprocessor chip (GOLD-SX) can be used.

The integration of the HiGOLD chipset in a GSM handy is presented in Figure C.1.

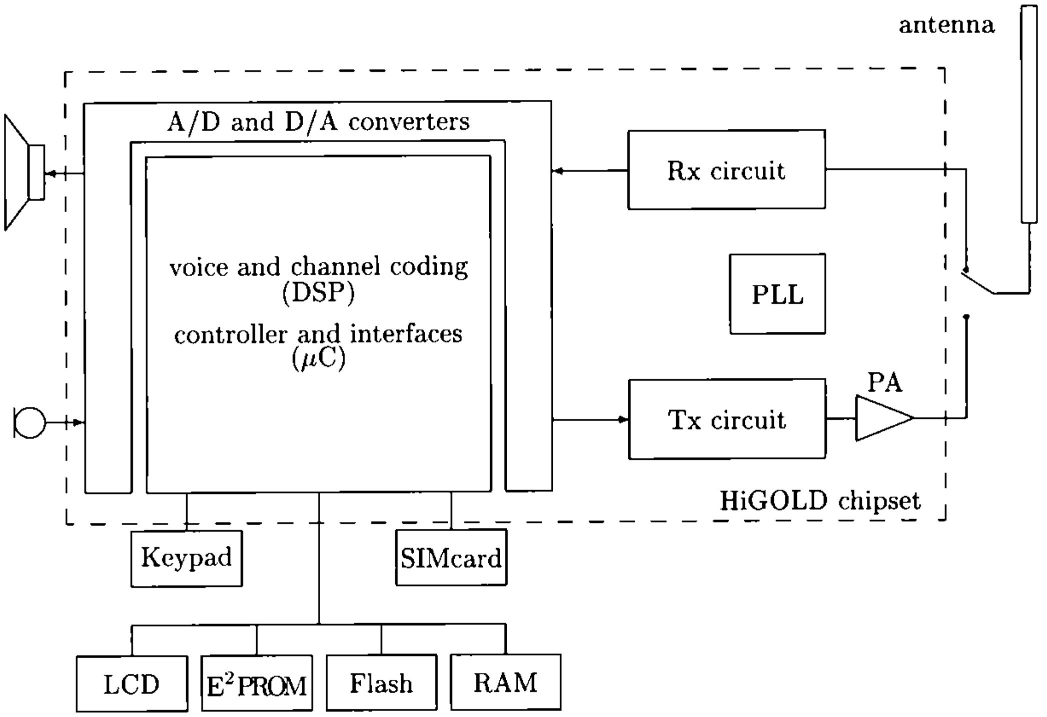


Figure C.1: HiGOLD system integration

The signal received from the antenna first passes a low noise amplifier, part of the RF demodulator. After external filtering, the RF signal is downconverted to an *Intermediate Frequency (IF)* by a first mixer stage of the RF receiver. An external *Surface Acoustic Wave (SAW)* filter performs a rough channel selection. The IF signal will be demodulated to baseband by a second mixer, after a previous digitally programmable gain-controlled amplification.
The resulting differential baseband signal is fed to the receive path of the GAIM, where both components, I and Q, are converted independently from each other into the digital domain.
Signal reconstruction and filtering of the digital baseband signal is performed in the signal processing part of the HiGOLD. A complex equalizer with soft-output recovers the original data stream.
In the case of GSM fullrate operation, data processing is continued on the digital signal processing part of the HiGOLD (Figure C.2) with soft-decision channel decoding and speech decoding, including comfort noise generation during discontinuous reception. In the case of halfrate or enhanced fullrate operation, the corresponding soft-decision channel decoder is part of the HiGOLD whereas the speech decoder (including DRX) is part of the coprocessor circuit *GOLD-SX*.
After voiceband interpolation on HiGOLD, the resulting data stream is digital-to-

analog converted and amplified by a programmable gain stage in the voiceband processing part of the GAIM. The output signal can be directly connected to a handset earpiece.

In the opposite direction the GAIM will amplify the input signal from the microphone. The amplifier is gain programmable. After analog-to-digital conversion the data stream is forwarded to the HiGOLD for voiceband decimation. In the case of GSM fullrate operation, data processing is continued with speech encoding including VAD and DTX. Channel encoding is followed by digital GMSK modulation. In the case of halfrate or enhanced fullrate operation, speech encoding (including VAD and DTX) is performed on the coprocessor circuit.
After modulation, the 10-bit I and Q baseband components are delivered to the baseband processing part of the GAIM where they are digital-to-analog converted. The resulting analog differential baseband signal is fed to the input of the RF modulator circuit. Here a quadrature amplitude modulator (QAM) directly converts the baseband to radio frequency (900 MHz or 1,800 MHz respectively). Finally an RF power module amplifies the RF signal to the required power. The ramping of the power amplifier is controlled by the system interface functions of the HiGOLD. The control values according to the prescribed ramping curves are digital-to-analog converted by the GAIM and passed on to the power amplifier.

From the digital signal processing point of view, the HiGOLD (Figure C.2) and the GAIM are the most important chips and will therefore be referred to in more detail.

The microcontroller part of the HiGOLD contains a 16-bit microcontroller and a system interface block which comprises a series of GSM-specific interfaces and control functions such as system interface with RF synthesizer, *Automatic Gain Control (AGC)*, *Automatic Frequency Control (AFC)* and *Power Amplification (PA)* control, chipcard interfacing, timing signal generation, clock generation. The digital signal processing part of the HiGOLD consists of two signal processing cores each with 26 MIPS and all the program and data memory required for fullrate operation, halfrate channel encoding and decoding and enhanced fullrate channel encoding and decoding for speech and data.
The HiGOLD-SP contains a fullrate speech codec (RPE-LTP), a channel codec with soft-decision decoding (bit-by-bit) and a complex soft-output (Viterbi) equalizer, frequency correction burst handling, all as DSP firmware. Moreover, signal processing dedicated hardware performs

- digital decimation for the received baseband signals
- digital interpolation for the received voiceband signals and digital decimation for the voiceband signals to be transmitted

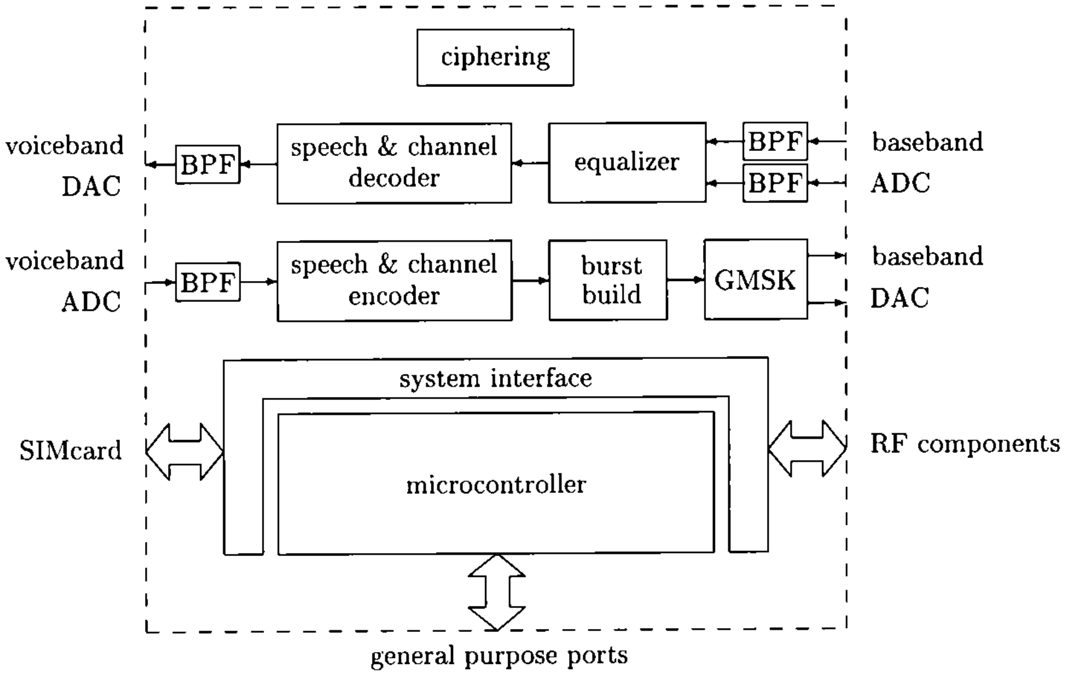


Figure C.2: HiGOLD block diagram

- cipher sequence generation according to the A51 and A52 algorithms
- burst generation, serial encryption and GMSK modulation

Analog-to-digital and digital-to-analog conversion of baseband and voiceband signals is performed in the GAIM. Furthermore the digital-to-analog conversion of an RF power control signal will be performed on this chip.

As to the GSM software, it is organized in GSM Layer 1, Protocol Stack (Layer 2 and 3) and the *Man Machine Interface (MMI)*. Software updates can be performed by reloading the Flash.

Appendix D

Composite Source Signal

The basic idea behind the *Composite Source Signal (CSS)* is to provide a test signal having both the typical characteristics of real speech and short duration for the measurement of short-term characteristics. The CSS yields good agreement with real speech when used for measurement of convergence characteristics of acoustic echo cancellers [Gilloire 94] and for simulation of double-talk periods. In Figure D.1 the standardized composite source signal from [ITU-T P.501 96] is presented.

The CSS is composed of three segments repeated sequentially:

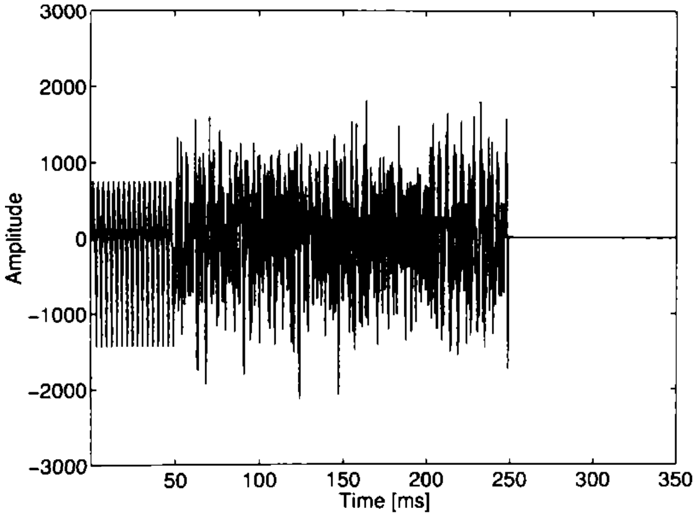


Figure D.1: Composite source signal

- a voiced signal of approximately 50 ms duration, used to activate speech detectors. The voiced signal can be described by a sequence of 134 16-bit words. According to a sampling rate of 44.1 kHz, this sequence will be repeated 16 times to achieve a duration of approximately 50 ms.

- a deterministic signal with broad spectrum of 200 ms duration, used for measuring short-term transfer functions. This signal has noise like features and is therefore called pseudo-noise signal. It is produced by specifying a complex spectrum with a constant magnitude and randomly changing phase. This spectrum will be inverse transformed by an Inverse Fast Fourier Transform producing the time signal.

- a pause of 100 - 150 ms duration

When using the CSS for measurements, the sequence of voiced sound, pseudo-noise signal and pause can be cycled. This means that after the pause, the sequence of voiced sound is repeated. With this procedure sequences of any length can be produced.

There also exist standardized bandlimited (between 200 Hz and 3.6 kHz) composite source signals with speech-like power density spectrum, which can be used for the measurement of acoustic echo cancellers. Two sequences are defined, one for single talk and another for double-talk, their power density spectra being presented in Figure D.2 and Figure D.3.

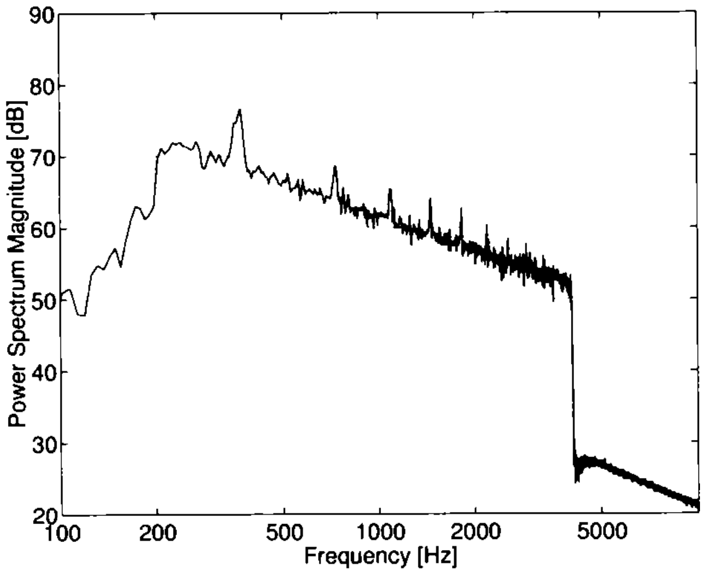


Figure D.2: Power density spectrum of single talk signal

Related to the single talk signal, the double-talk sequence has slightly different length of the voiced signal (approximately 75 ms) and the pause (approximately 125 ms). The voiced signal for double-talk also presents a different pitch frequency than the single talk voiced signal. Instead of the pseudo-noise of the single talk signal a white gaussian random noise signal is used. The total length of the double-talk signal is 400 ms.

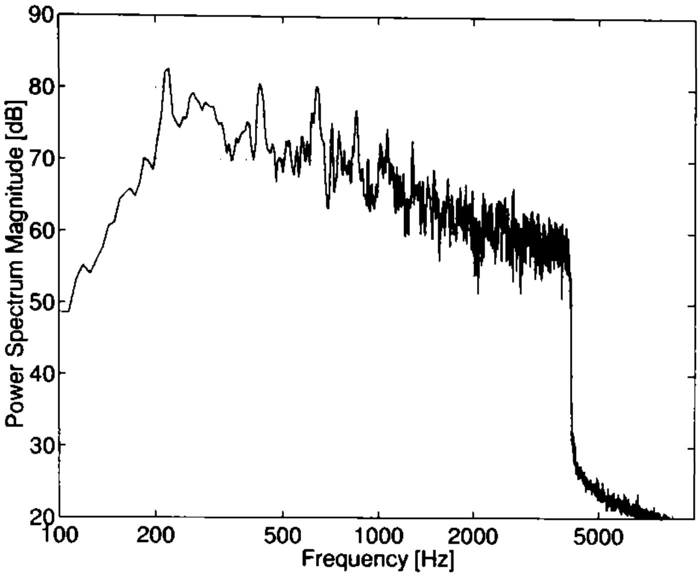


Figure D.3: Power density spectrum of double-talk signal

In this way a typical double-talk condition can be simulated, with the composite source single talk signal used as the far-end speech and the composite source double-talk signal employed as near-end speech [ITU-T P.501 96]. The correlation between single talk and double-talk will be low.
The double-talk condition with two signals applied simultaneously can be thus reproduced very realistically.

Appendix E

Singular Value Decomposition

The *Singular value decomposition* or *SVD* is a very powerful set of techniques for dealing with equations or matrices that are singular or very close to singular [Press et al. 92]. SVD methods are based on the following theorem of linear algebra:

Any $(M \times N)$ matrix $\mathbf{X}$ with $M \geq N$ can be written as the product of an $(M \times N)$ column-orthogonal matrix $\mathbf{U}$, an $(N \times N)$ diagonal matrix $\mathbf{W}$ with positive and zero elements (the singular values), and the transpose of an $(N \times N)$ orthogonal matrix $\mathbf{V}$.

$$\mathbf{X} = \mathbf{U}\mathbf{W}\mathbf{V}^T \tag{E.1}$$

The matrices $\mathbf{U}$ and $\mathbf{V}$ are each orthogonal in the sense that their columns are orthonormal, i.e. $\mathbf{U}^{-1} = \mathbf{U}^T$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ where $\mathbf{I}$ represents the identity matrix. Similar for $\mathbf{V}$). $\mathbf{X}^T$, the transpose of $\mathbf{X}$ can be written as

$$\mathbf{X}^T = \mathbf{V}\mathbf{W}^T\mathbf{U}^{=}\mathbf{V}\mathbf{W}\mathbf{U}^T \tag{E.2}$$

Now considering the product $\mathbf{X}^T\mathbf{X}$

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{W}\mathbf{U}^T\mathbf{U}\mathbf{W}\mathbf{V}^T = \mathbf{V}\mathbf{W}^2\mathbf{V}^T \tag{E.3}$$

the inverse of the matrix product will be

$$(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{V}\mathbf{W}^2\mathbf{V}^T)^{-1} = \mathbf{V}\mathbf{W}^{-2}\mathbf{V}^T \tag{E.4}$$

The last step in Eq. (E.4) is possible because both $\mathbf{V}$ and $\mathbf{W}$ are square and only for square matrices $\mathbf{\Phi}, \mathbf{\Psi}$ it can be written [Bronstein et al. 95]: $(\mathbf{\Phi}\mathbf{\Psi})^{-1} = \mathbf{\Psi}^{-1}\mathbf{\Phi}^{-1}$.

Thus, the computation of the adjustment vector in the conventional affine projection algorithm

$$\Delta\hat{\mathbf{h}}[k] = \mathbf{X}_p[k] \left(\mathbf{X}_p^T[k]\,\mathbf{X}_p[k]\right)^{-1} \mathbf{e}_p[k] \tag{E.5}$$

can be simplified by inverting just one diagonal matrix [Ansahl 98]:

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{U}\mathbf{W}\mathbf{V}^T\mathbf{V}\mathbf{W}^{-2}\mathbf{V}^T = \mathbf{U}\mathbf{W}^{-1}\mathbf{V}^T \tag{E.6}$$

The SVD can also be carried out when $M \leq N$, in this case the singular values w_j for $j = M+1, \ldots, N$ will be all zero and the corresponding columns of $\mathbf{U}$ will also be zero. The decomposition (E.1) can always be done, no matter how singular the matrix $\mathbf{X}$ is.

Appendix F

Impulse Response Measurement in Cars

The impulse response as the most fundamental physical descriptor in room acoustics is determined by studying the response of the system to a particular excitation. All the acoustical measures for evaluations of rooms can be derived from the impulse response. One of the most powerful methods involves the use of pseudo-random sequences as a source of excitation [Otshudi et al. 88]. A *pseudo-random sequence (PRS)*, also known as *maximum-length sequence* or *m-sequence*, is a periodic binary sequence that has approximately flat spectrum which can be generated quite simply by a feedback arrangement of shift registers [Chu 90]. The most important properties of the pseudo-random sequence are:

- it is periodic with length $n = 2^m - 1$ where m is the number of stages used in the shift register arrangement
- the periodic autocorrelation function [MacWilliams & Sloane 76] is given by
$$r[k] = \begin{cases} 1 & \text{for} \quad k = 0 \\ -\dfrac{1}{n} & \text{for} \quad 1 \leq k \leq n-1 \end{cases} \tag{F.1}$$
The corresponding power spectrum is a line spectrum. For linear time invariant systems one period of the signal is sufficient and no averaging will be required [Chu 87].
- if a window of length m is slid along the pseudo-random sequence, each of the $2^m - 1$ non-zero binary m-tuples will be seen exactly once.

To construct a pseudo-random sequence of length $n = 2^m - 1$ a primitive polynomial $h(x)$ of degree m is needed. This polynomial specifies a feedback shift register consisting of m binary memory elements. At each time instant the contents of the memory elements is shifted one place to the right and the elements corresponding to the terms in $h(x)$ are added modulo-2 and fed back into the left-hand element. In Figure F.1 the feedback shift register corresponding to the primitive polynomial

$$h(x) = x^4 + x + 1 \tag{F.2}$$

is presented. The symbol $\oplus$ denotes a modulo-2 addition. The period of the primitive polynomial from Eq. (F.2) is $n = 2^4 - 1 = 15$ and, if an initial state of 1000 is considered the output sequence from the shift register will be:

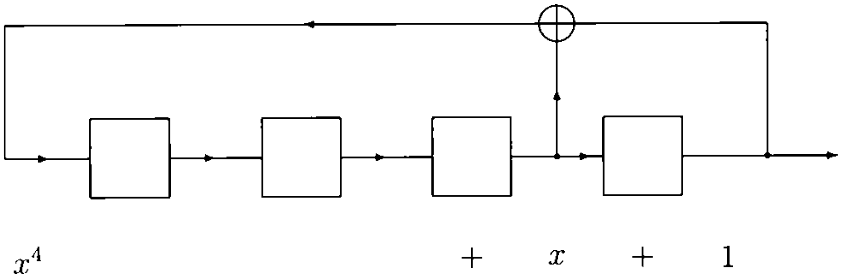


Figure F.1: Feedback shift register example for Pseudo-random sequence generation

```
0 0 0 1 0 0 1 1 0 1 0 1 1 1 1
0 0 1 0 0 1 1 0 1 0 1 1 1 1 0
0 1 0 0 1 1 0 1 0 1 1 1 1 0 0
1 0 0 1 1 0 1 0 1 1 1 1 0 0 0
0 0 1 1 0 1 0 1 1 1 1 0 0 0 1
0 1 1 0 1 0 1 1 1 1 0 0 0 1 0
1 1 0 1 0 1 1 1 1 0 0 0 1 0 0
1 0 1 0 1 1 1 1 0 0 0 1 0 0 1
0 1 0 1 1 1 1 0 0 0 1 0 0 1 1
1 0 1 1 1 1 0 0 0 1 0 0 1 1 0
0 1 1 1 1 0 0 0 1 0 0 1 1 0 1
1 1 1 1 0 0 0 1 0 0 1 1 0 1 0
1 1 1 0 0 0 1 0 0 1 1 0 1 0 1
1 1 0 0 0 1 0 0 1 1 0 1 0 1 1
1 0 0 0 1 0 0 1 1 0 1 0 1 1 1
```

The use of pseudo-random sequences in the impulse response determination of a room is based on the fact that the input-output crosscorrelation of a linear time-invariant system under white noise excitation is proportional to the system's impulse response [Chu 90]. Thus, the input signal $S_i(t)$ and the output signal $S_o(t)$ are related through the crosscorrelation

$$h(\tau) = R_{io}(\tau) = \frac{1}{T}\int_0^T S_i(t-\tau)S_o(t)dt \tag{F.3}$$

for the signals represented in Figure F.2

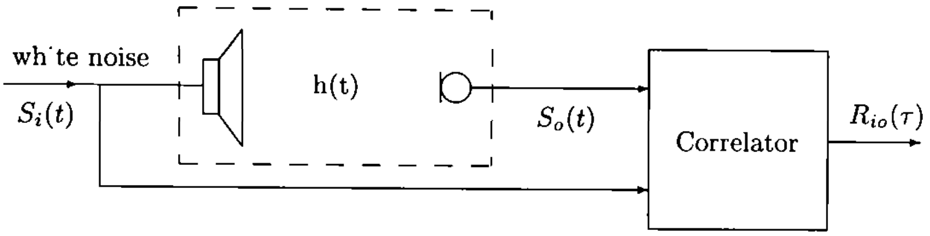


Figure F.2: Impulse response measurement block diagram

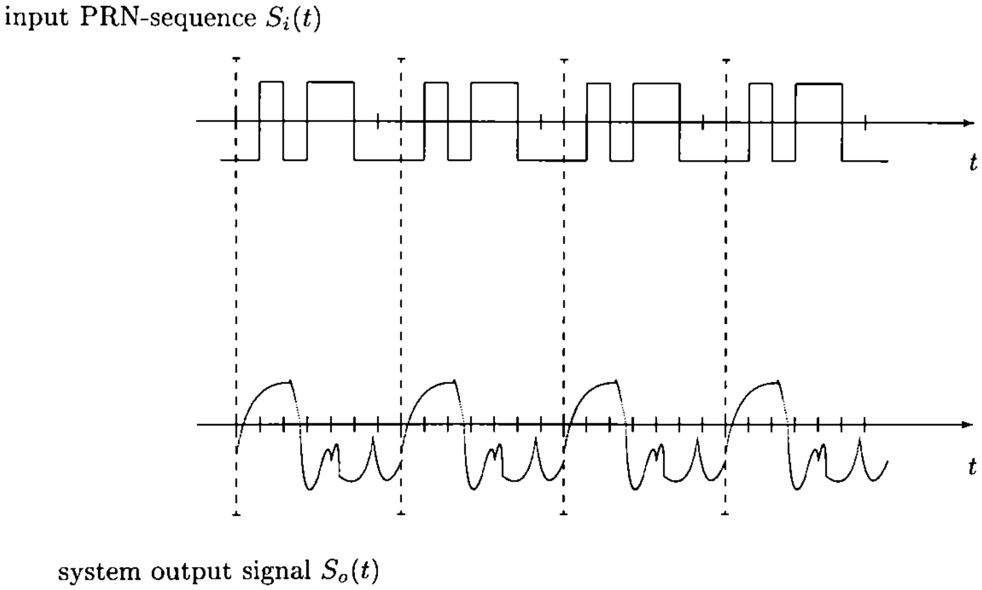


Figure F.3: Example of m-sequence input signal and a hypothetical output signal

As the pseudo-random noise has a flat spectrum, it can be considered as white noise. For signal processing, the binary states of 0 and 1 of the m-sequence will be changed to +1 and −1. In Figure F.3 an example is shown of an m-sequence of length $n = 7$ as the input signal to the linear time-invariant system $S_i(t)$ and $S_o(t)$ is considered a hypothetical output. If $S_o(t)$ is sampled at the clock frequency of the pseudo-random sequence, Eq. (F.4) can be expressed in matrix form

$$\mathbf{h} = \frac{1}{n} \mathbf{M} \, \mathbf{S_o} \tag{F.4}$$

where **h** is the impulse response vector of length n, **M** represents the $(n \times n)$ matrix containing the right circularly delayed version of the m-sequence and $\mathbf{S_o}$ represents the output signal vector of length n.

As the elements of **M** are only +1 or −1, a fast computation of the product

$\mathbf{M}$ $\mathbf{S}_o$ is possible by applying the techniques developed in Hadamard spectroscopy [Chu 87]. A Hadamard's matrix is a square matrix of dimension 2^n, consisting of elements +1 and −1, whose rows or columns are mutually orthogonal [Otshudi et al. 88]. The matrix **M** can be transformed into a Hadamard matrix by adding a row and a column of +1 and followed by a reordering of the rows and columns. The matrix multiplication will be performed in five steps [Chu 90]:

- matrix **M** can be factored into two matrices: **R** $(n \times m)$ and **C** $(m \times n)$. **R** is obtained by choosing those columns of **M** such that the first m rows of **R** form a $(m \times m)$ unit matrix. **C** is contains the first m rows of **M**.

- the row tags of **R** and the column tags of **C** are obtained according to the integer equivalence of their m-bit binary digits.

- the columns of **M** will be reordered by using the tags of **C** while the rows will be reordered by using the tags of **R**. Furthermore a row and a column of +1 will be added to **M**, thus generating the Hadamard matrix **H**.

- the elements of vector S_o will be reordered following the column tags of **C** and a zero element will be added as first element.

- the resulting vector **h** will have its first element omitted and the following elements will be reordered using the tags of **R**

This technique is independent of any possibly existing background noise, because there is no correlation between the pseudo-random sequence and any other background noise [Chu 87].

List of Figures

List of Tables

Bibliography

[Ahmed 89] Ahmed, M.S.: "Comparison of noisy speech enhancement algorithms in terms of LPC perturbation", *IEEE Trans. on ASSP*, vol. ASSP-37, no. 1, pp. 121-125, Jan. 1989

[Al-Hashemy & Taha 88] Al-Hashemy, B.A.R., Taha, S.M.R.: "Voiced-unvoiced-silence classification of speech signals based on statistical approaches", *Applied Acoustics*, 25, pp. 169-179, 1988

[Ansahl 98] Ansahl, T.: *Untersuchung eines Echokompensationsalgorithmus für Mobiltelefon-Freisprecheinrichtungen im Auto*, Diploma Thesis, Technische Universität München, Jan. 1998

[Ansahl et al. 98] Ansahl, T., Varga, I., Kremmer, I. and Xu, W.: "Adaptive acoustic echo cancellation based on FIR and IIR filter banks", Paper proposal no. 2283 for *ICASSP 99*, Phoenix, Arizona, USA, March 1999

[Armbruster et al. 91] Armbruster, W., Dobler, S. and Meyer, P.: "Hands-free telephony, speech recognition and speech coding techniques implemented in the SPS51", *Philips Telecommunication Review*, vol. 49, no. 1, pp. 19-27, March 1991

[Ayad & Faucon 95] Ayad, B. and Faucon, G.: "Acoustic echo and noise cancellation for hands-free communication systems", *4-th Int. Workshop on Acoustic Echo and Noise Control*, Trondheim, Norway, Norwegian Inst. Technol., pp. 91-94, June 1995

[Ayad et al. 96] Ayad, B., Faucon, G. and Le Bouquin–Jeannès, R.: "Optimization of a noise reduction preprocessing in an acoustic echo and noise controller", *Proc. of ICASSP 96*, Atlanta, GA, USA, vol. 2, pp. 953-956, May 1996

[Baillargeat et al. 91] Baillargeat, C., Boudy, J. and Lockwood, P.: "Noise reduction for speech enhancement in cars: non-linear spectral subtraction / Kalman filtering", *Proc. of EUROSPEECH '91*, pp. 83-86, 1991

[Berouti et al. 79] Berouti, M., Schwartz, R. and Makhoul, J.: "Enhancement of speech corrupted by acoustic noise", *Proc. of IEEE Conf. on ASSP*, pp. 208-211, April 1979

[Boll 79] Boll, S.: "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. on ASSP*, vol. ASSP-27, no. 2, pp. 113-120, 1979

[Boudy et al. 95] Boudy, J., Capman, F. and Lockwood, P.: "A globally optimised frequency-domain acoustic echo canceller for adverse environment applications", *4-th Int. Workshop on Acoustic Echo and Noise Control*, Trondheim, Norway, Norwegian Inst. Technol., pp. 95-98, June 1995

[Bronstein et al. 95] Bronstein, I. L., Semendjajew, K. A., Musiol, G. and Mühlig, G.: *Taschenbuch der Mathematik*, 2., überarbeitete und erweiterte Auflage, Thun, Frankfurt am Main: Verlag Harri Deutsch, 1995

[Campbell 93] Campbell, D.R.: "Speech enhancement and the automotive environment", *ISATA-Proc.: Mechatronics*, pp. 51-58, 1993

[Carter 87] Carter, G.: "Coherence and time delay estimation", *Proc. IEEE*, vol. 75, no. 2, pp. 236-255, Feb. 1987

[connect 98] "Anschlußzahlen des Monats", *connect*, no. 12, p. 14, Dec. 98

[Ching et al. 93] Ching, W.S., Toh, P.S. and Yuan Baozong: "Robustness of signal correlation", *Proc. of TENCON '93, IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering*, vol. 3, pp. 279-282, Beijing, China, Oct. 1993

[Chu 87] Chu, W. T.: "A deterministic broad-band signal for acoustical measurements", *Inter-Noise 87*, pp. 1199-1202, 1987

[Chu 90] Chu, W. T.: "Impulse-response and reverberation-decay measurements made by using a periodic pseudorandom sequence", *Applied Acoustics*, 29, pp. 193-205, 1990

[Chu & Messerschmitt 82] Chu, P.L. and Messerschmitt, D.G.: "A weighted Itakura-Saito spectral distance measure", *IEEE Trans. on ASSP*, vol. ASSP-30, no. 4, pp. 545-560, Aug. 1982

[Crochiere & Rabiner 83] Crochiere, R. E. and Rabiner, L. R.: *Multirate Digital Signal Processing*, Englewood Cliffs, New Jersey: Prentice-Hall, 1983

[Crozier et al. 93] Crozier, P.M., Cheetham, B.M.G., Holt, C. and Munday, E.: "The use of linear prediction and spectral scaling for improving speech enhancement", *Proc. of EUROSPEECH '93*, Berlin, pp. 231-234, Sept. 1993

[Curtis & Niederjohn 78] Curtis, R.A. and Niederjohn, R.J.: "An investigation of several frequency-domain processing methods for enhancing the intelligibility of speech in wideband random noise", *Proc. of ICASSP 78*, Tulsa, USA, pp. 602-605, April 1978

[Dal Degan & Prati 88] Dal Degan, N. and Prati, C.: "Acoustic noise analysis and speech enhancement techniques for mobile radio applications", *Signal Processing*, vol. 15, no. 1, pp. 43-56, July 1988

[DeGroat et al. 97] DeGroat, R.D., Begusic, D., Dowling, E.M. and Linebarger, D.A.: "Spherical subspace and eigen based affine projection algorithms", *Proc. of ICASSP 97*, pp. 2345-2348, Munich, Germany, April 1997

[Deller et al. 93] Deller, J. R., Proakis, J. G. and Hansen, J. H. L.: *Discrete-Time Processing of Speech Signals*, Upper Saddle River, NJ: Prentice Hall, 1993

[El-Maleh & Kabal 97] El-Maleh, K. and Kabal, P.: "Comparison of voice activity detection algorithms for wireless personal communication systems", *Canadian Conference on Electrical and Computer Engineering*, vol. 2, pp. 470-473, May 1997

[Faucon & Le Bouquin-Jeannès 95] Faucon, G. and Le Bouquin-Jeannès, R.: "Joint system for acoustic echo cancellation and noise reduction", *Proc. of EUROSPEECH '95*, Madrid, pp. 1525-1528, Sept. 1995

[Faucon & Tazi Mezalek 90] Faucon, G. and Tazi Mezalek, S.: "Theoretical comparison of two noise reduction methods", *Signal Processing V: Theories and Applications*, pp. 1963-1966, 1990

[Faucon et al. 89] Faucon, G., Tazi Mezalek, S. and Le Bouquin, R.: "Study and comparison of three structures for enhancement of noisy speech", *Proc. of ICASSP 89*, pp. 385-388, 1989

[Ferrara & Widrow 81] Ferrara, E.R. and Widrow, B.: "Multichannel adaptive filtering for signal enhancement", *IEEE Trans. on ASSP*, vol. 29, no. 3, pp. 766-770, 1981

[Fliege 93] Fliege, N.: *Multiraten-Signalverarbeitung - Theorie und Anwendungen*, Stuttgart, Germany: Teubner, 1993

[Freeman et al. 89] Freeman, D.K., Cosier, G. Southcott, C.B. and Boyd, I.: "The voice activity detector for the pan-European digital cellular mobile telephone service", *Proc. of ICASSP 89*, vol. 1, S7.6, pp. 369-372, 1989

[Frenzel 92] Frenzel, R.: *Freisprechen in gestörter Umgebung*, Doctoral Dissertation, Technische Hochschule Darmstadt, in *Fortschrittsberichte, VDI Reihe 10: Informatik und Kommunikationstechnik*, Nr. 228, Düsseldorf: VDI-Verlag, 1992

[Gänsler et al. 96] Gänsler, T., Hansson, M., Ivarsson, C.-J. and Salomonsson, G.: "A double-talk detector based on coherence", *IEEE Trans. on Communications*, vol. 44, no. 11, pp. 1421-1427, Nov. 1996

[Ganapathiraju & Picone 97] Ganapathiraju, A. and Picone, J.: "Echo cancellation for evaluating speaker identification technology", *Proc. of IEEE SOUTHEASTCON '97*, Blacksburg, USA, pp. 100-102, April 1997

[Gay & Tavathia 95] Gay, S. L. and Tavathia, S.: "The fast affine projection algorithm", *Proc. of ICASSP 95*, Detroit, MI, USA, pp. 3023-3026, May 1995

[Gilloire 94] Gilloire, A.: "Performance evaluation of acoustic echo control: required values and measurement procedures", *Annales des Télécommunication*, vol. 49, no. 7-8, pp. 368-372, July-Aug. 1994

[Gilloire 95] Gilloire, A.: "Recent advances in adaptive filtering algorithms for acoustic echo cancellation" *4-th Int. Workshop on Acoustic Echo and Noise Control*, Trondheim, Norway, Norwegian Inst. Technol., pp. 115-134, June 1995

[Gilloire & Vetterli 92] Gilloire, A. and Vetterli, M.: "Adaptive filtering in subbands with critical sampling: analysis, experiments and application to acoustic echo cancellation", *IEEE Trans. on Signal Processing*, vol. 40, no. 8, pp. 1862-1875, Aug. 1992

[Goldenberg & Bisson 95] Goldenberg, O. and Bisson, F.: "Improving GSM service quality: the new solutions provided by the echo canceller", *Communication & Transmission*, vol. 17, no. 3, pp. 74-82, 1995

[Goubran et al. 90] Goubran, R.A., Hebert, R. and Hafez, H.M.: "Acoustic noise suppression using regressive adaptive filtering", *IEEE Trans. on Vehicular Technology Conference*, pp. 48-53, 1990

[Goulding & Bird 90] Goulding, M.M. and Bird, J.S.: "Speech enhancement for mobile telephony", *IEEE Trans. Vehicular on Technology Conference*, vol. 39, no. 4, pp. 316-326, Nov. 1990

[Gray & Markel 76] Gray, Jr. A. and Markel, J.D.: "Distance measures for speech processing", *IEEE Trans. on ASSP*, vol. ASSP-24, no. 5, pp. 380-391, Oct. 1976

[GSM Data 98] GSM Data Knowledge Site: "1998 Subscriber unit sales and subscriber forecast", *http://www.gsmdata.com/subscriber.htm*, Nov. 10, 1998

[GSM MoU 98a] GSM Memorandum of Understanding Association: "Membership Statistic", *http://www.gsmworld.com/assoc/stats.htm*, Nov. 10, 1998

[GSM MoU 98b] GSM MoU: "The GSM Memorandum of Understanding - How it works", *http://www.cellular.co.za/gsm-mou.htm*, Nov. 12, 1998

[GSM Rec. 03.50 96] GSM Recommendation 03.50: "Digital cellular telecommunications system (Phase 2+); Transmission planning aspects of the speech service in the GSM Public Land Mobile Network (PLMN) system", Version 5.0.2, April 1997

[GSM Rec. 05.03 95] GSM Recommendation 05.03: "Digital cellular telecommunications system (Phase 2+); Channel coding", Version 5.2.0, Aug. 1996

[GSM Rec. 06.10 95] GSM Recommendation 06.10: "European digital cellular telecommunications system (Phase 1); GSM full rate speech transcoding", Version 3.2.0, Jan. 1995

[GSM Rec. 06.32 95] GSM Recommendation 06.32: "European digital cellular telecommunications system (Phase 1); Voice activity detection", Version 3.0.0, Jan. 1995

[GTR SMG 97] GSM Technical Report, Draft GTR/SMG: "Characteristics and test methods and quality assessment for handsfree mobile stations", April 1997

[Guelou et al. 96] Guelou, Y., Benamar, A. and Scalart, P.: "Analysis of two structures for combined acoustic echo cancellation and noise reduction", *Proc. of ICASSP 96*, Atlanta, GA, USA, vol.2, pp. 637-640, May 1996

[Gustafsson et al. 96] Gustafsson, S., Martin, R. and Vary, P.: "On the optimization of speech enhancement systems using instrumental measures", *EURASIP 96*, Darmstadt, Germany, pp. 36-40, March 1996

[Häkkinen & Väänänen 93] Häkkinen, J. and Väänänen, M.: "Background noise suppressor for a car hands-free microphone", *Proc. of the 4-th Int. Conf. on Signal Processing Applications and Technology*, vol. 1, pp. 300-307, Santa Clara, USA, Sept. 1993

[Harrison et al. 86] Harrison, W.A., Lim, J.S. and Singer, E.: "A new application of adaptive noise cancellation", *IEEE Trans. on ASSP*, vol. ASSP-34, no. 1, pp. 21-27, Feb. 1986

[Haykin 96] Haykin, S.: *Adaptive Filter Theory*, Third edition, Upper Saddle River, NJ: Prentice Hall, 1996

[Heitkämper 94] Heitkämper, P.: "Ein Korrelationsmaß zur Feststellung von Sprecheraktivitäten", *Proc. of 8. Aachener Kolloquium Signaltheorie*, pp. 97-100, Aachen, Germany, 1994

[Heitkämper 97] Heitkämper, P.: "An adaptation control for acoustic echo cancellation", *IEEE Signal Process. Letters*, vol. 4, no. 6, pp. 170-172, USA, June 1997

[Heitkämper & Walker 93] Heitkämper, P. and Walker, M.: "Adaptive gain control and echo cancellation for hands-free telephone systems", *Proc. of the 3-rd European Conf. on Speech Communication and Technology*, pp. 1077-1080, Berlin, 1993

[Hirano & Sugiyama 95] Hirano, A. and Sugiyama, A.: "A noise-robust stochastic gradient algorithm with an adaptive step-size suitable for mobile hands-free telephones", *Proc. of ICASSP 95*, Detroit, MI, USA, pp. 1392-1395, May 1995

[Hoyt & Wechsler 94] Hoyt, J.D. and Wechsler, H.: "Detection of human speech in structured noise", *Proc. of ICASSP 94*, Australia, pp. II-237-II-240, May 1994

[ITU Rec. G.729] ITU-T Recommendation G.729 - Annex B: "A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70", Nov. 1996

[ITU-T G.131 96] ITU-T Recommendation G.131 (08/96): "Stability and echo", 1996

[ITU-T G.165 93] ITU-T Recommendation G.165 (03/93): "Echo cancellers", 1993

[ITU-T G.167 93] ITU-T Recommendation G.167 (03/93): "Acoustic Echo Controllers", 1993

[ITU-T P.501 96] ITU-T Recommendation P.501 (08/96): "Test signals for use in telephonometry", 1996

[Johnson et al. 90] Johnson Jr., C.R., Ding, Z. and Sethares, W.A.: "Frequency-dependent bursting in adaptive echo cancellation and its prevention using double-talk detectors", *Int. Journal Adapt. Cont. Signal Processing*, vol. 4, pp. 219-236, 1990

[Kaneda et al. 95] Kaneda, Y., Tanaka, M. and Kojima, J.: "An adaptive algorithm with fast convergence for multi-input sound control", *ACTIVE 95*, Newport Beach, CA, USA, pp. 993-1004, July 1995

[Klema & Laub 80] Klema, V.C. and Laub, A.J.: "The singular value decomposition: Its computation and some applications", *IEEE Trans. Autom. Control*, vol. AC-25, pp. 164-176, 1980

[Kobatake et al. 78] Kobatake, H., Inari, J. and Kakuta, S.: "Linear predictive coding of speech signals in a high ambient noise environment", *Proc. of ICASSP 78*, pp. 472-475, June 1978

[Kremmer 98] Kremmer, I.: "Stepsize control of the affine projection algorithm in noisy environment", submitted to *EUROSPEECH '99*, Budapest, Sept. 1999

[Kremmer & Ansahl 98] Kremmer, I. and Ansahl, T.: "Acoustic echo cancellation in noisy environment using the affine projection algorithm in subbands", submitted to *Signal Processing Advances in Wireless Communication*, Annapolis, MD, USA, May 1999

[Kuo & Pan 93] Kuo, S.M. and Pan, Z.: "Distributed acoustic echo cancellation system with double-talk detector", *Journal of the Acoustical Society of America*, vol. 94, no. 6, pp. 3057-3060, Dec. 1993

[Kuo & Pan 94] Kuo, S.M. and Pan, Z: "Development and analysis of distributed acoustic echo cancellation microphone system", *Signal Processing*, vol. 37, no. 3, pp. 333-344, June 1994

[Kuttruff 91] Kuttruff, H.: *Room Acoustics*, Third edition, London: Applied Science Publishers Ltd., 1991

[Le Bouquin & Faucon 90] Le Bouquin, R. and Faucon, G.: "On using the coherence function for noise reduction", *Signal Processing V: Theories and Applications*, pp. 1103-1106, 1990

[Le Bouquin & Faucon 92] Le Bouquin, R. and Faucon, G.: "Study of a noise cancellation system based on the coherence function", *Proc. of EUSIPCO-92*, Brussels, Belgium, vol. 3, pp. 1633-1636, Aug. 1992

[Le Bouquin et al. 93] Le Bouquin, R., Faucon, G. and Akbari Azirani, A: "Proposal of a composite measure for the evaluation of noise cancelling methods in speech processing", *Proc. of EUROSPEECH '93*, Berlin, pp. 227-230, Sept. 1993

[Le Bouquin-Jeannès et al. 94] Le Bouquin-Jeannès, R., Faucon, G., Akbari Azirani, A. and Ehrmann, F.: "Speech enhancement using sub-band decomposition and comparison with full-band techniques", *Proc. of EUSIPCO-94*, Edinburgh, UK, vol.2, pp. 1206-1209, Sept. 1994

[Le Bouquin-Jeannès et al. 96] Le Bouquin-Jeannès, R., Faucon, G. and Ayad, B.: "How to improve acoustic echo and noise cancelling using a single talk detector", *Speech Communication (Netherlands)*, vol. 20, no. 3-4, pp. 191-202, Dec. 1996

[Liberti et al. 91] Liberti, J.C., Rappaport, T.S. and Proakis, J.G.: "Evaluation of several adaptive algorithms for canceling acoustic noise in mobile radio environments", *IEEE Trans. on Vehicular Technology Conference*, St. Louis, USA, pp. 126-132, May 1991

[Lim & Oppenheim 78] Lim, J.S. and Oppenheim, A.V.: "All-pole modelling of degraded speech", *IEEE Trans. on ASSP*, vol. ASSP-26, no. 3, pp. 197-210, June 1978

[Lim & Oppenheim 79] Lim, J.S. and Oppenheim, A.V.: "Enhancement and bandwidth compression of noisy speech", *Proc. of IEEE*, vol. 67, no. 12, pp. 1586-1604, Dec. 1979

[Lim & Wang 82] Lim, J.S. and Wang, D.Y.: "The unimportance of phase in speech enhancement", *IEEE Trans. on ASSP*, vol. ASSP-30, pp. 679-681, 1982

[Lockwood et al. 91] Lockwood, P., Baillargeat, C., Boudy. J. and LeLievre, L.: "Robust techniques for speech processing in car adverse environments for GSM applications", *6-th World Telecommunication Forum. Part 2. Technical Symposium. Integration, Interoperation and Interconnection: The Way to Global Services*, Geneva, Switzerland, vol. 1, pp. 327-331, Oct. 1991

[Lockwood & Boudy 92] Lockwood, P. and Boudy, J.: "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in car", *Speech Communications*, vol. 11, no. 2-3, pp. 215-218, June 1992

[Lynn & Fuerst 89] Lynn, P.A. and Fuerst, W.: *Introductory Digital Signal Processing with Computer Applications*, New York: John Wiley & Sons Ltd., 1989

[MacWilliams & Sloane 76] MacWilliams, F. J. and Sloane, N. J. A.: "Pseudorandom sequences and arrays", *Proc. of IEEE*, vol. 64, no. 12, pp. 1715-1730, Dec. 1976

[Marple 87] Marple, S.L. Jr.: *Digital Spectral Analysis with Applications*, Englewood Cliffs, New Jersey: Prentice-Hall, 1987

[Martin & Vary 94] Martin, R. and Vary, P.: "Combined acoustic echo cancellation, deverberation and noise reduction: a two microphone approach", *Annales des Télécommunication*, vol. 49, no. 7-8, pp. 429-438, July-Aug. 1994

[Martin et al. 96] Martin, R., Vary, P., Ramponi, G. and Marsi, S.: "Combined acoustic echo control and noise reduction for hands-free telephony - state of the art and perspectives", *Proc. of EUSIPCO-96*, Trieste, Italy, vol. 2, pp. 1107-1110, Sept. 1996

[McAlinden & Hartley 90] McAlinden, P. and Hartley, D.: "Vernetztes Europa", *Elektronik*, no. 25, pp. 46-52, 1990

[Meana et al. 94] Meana, H. P., de Rivera, L. N., Miyatake, M. N., Sanchez, F. C. and Garcia, J. C. S.: "A time varying step size normalized LMS echo canceller algorithm", *Proc. of ICASSP 94*, vol. 2, pp. 249-252, 1994

[Munday 88] Munday, E.: "Noise reduction using frequency-domain non-linear processing for the enhancement of speech", *British Telecom Technology Journal*, vol. 6, no. 2, pp. 71-83, April 1988

[Naylor et al. 94] Naylor, P., Alcazar, J., Boudy, J. and Grenier, Y.: "Enhancement of hands-free telecommunications", *Annales des Télécommunication*, vol. 49, no. 7-8, pp. 373-379, July-Aug. 1994

[O'Shaughnessy 89] O'Shaughnessy, D.: "Enhancing speech degraded by additive noise or interfering speakers", *IEEE Communications Magazine*, pp. 46-52, Feb. 1989

[Oh et al. 97] Oh, S., Linebarger, D., Priest, B. and Raghothaman, B.: "A fast affine projection algorithm for an acoustic echo canceller using a fixed-point DSP processor", *Proc. of ICASSP 97*, vol. 5, pp. 4121-4124, Munich, Germany, April 1997

[Otshudi et al. 88] Otshudi, L., Guilhot, J. P. and Charles, J. L.: "Overview of techniques for measuring impulse response in room acoustics", *Proc. of The Institute of Acoustics*, vol. 10, part 2, pp. 407-414, 1988

[Ozeki & Umeda 84] Ozeki, K. and Umeda, T.: "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties", *Electronics and Communications in Japan*, vol. 67-A, no. 5, pp. 19-27, 1984

[Pauler 98] Pauler, W.: "Mangelhafte Akustik. Freisprecheinrichtungen im Test", *Funkschau*, no. 7, pp. 26-28, 30-33, 1998

[Pollak et al. 93] Pollak, P., Sovka, P. and Uhlíř, J.: "Noise suppresion system for a car", *Proc. of EUROSPEECH '93*, Berlin, pp. 1073-1076, Sept. 1993

[Powell 94] Powell, D.: "Echo complexities in the digital domain", *Telephony (USA)*, vol. 227, no. 19, pp. 38-40, Nov. 1994

[Press et al. 92] Press, W. H., Teukolsky, S. A., Vetterling, W.T. and Flannery, B.P.: *Numerical Recipes in C: the Art of Scientific Computing*, Second edition, Cambridge: Cambridge University Press, 1992

[Quackenbush et al. 88] Quackenbush, S.R., Barnwell III, T.P. and Clements, M.A.: *Objective Measures of Speech Quality*, Englewood Cliffs, New Jersey: Prentice-Hall, 1988

[Rabiner & Sambur 75] Rabiner, L.R. and Sambur, M.R.: "An algorithm for determining the endpoints of isolated utterances", *Bell System Technical Journal*, vol.54, no.2, pp. 297-315, Feb. 1975

[Rabiner & Schafer 78] Rabiner, L. and Schafer, R.: *Digital Processing of Speech Signals*, Englewood Cliffs, New Jersey: Prentice-Hall, 1978

[Richardson & Gowdy 96] Richardson, J.B. and Gowdy, J.: "LPC-synthesis mixture: a low computational cost speech enhancement algorithm", *IEEE SOUTHEASTCON '96. Bringing Together Education, Science and Technology*, pp. 496-499, New York, 1996

[Rowden 91] Rowden, C.: "Analysis" in *Speech processing*, Rowden, C., Editor, London: McGraw Hill Book Company, 1991, pp. 35-73

[Rowden & Hall 91] Rowden, C. and Hall, S.: "Parametric coding of speech" in *Speech processing*, Rowden, C., Editor, London: McGraw Hill Book Company, 1991, pp. 158-183

[Sambur 79] Sambur, M.: "A preprocessing filter for enhancing LPC analysis/synthesis of speech", *Proc. of ICASSP 79*, pp. 971-974, 1979

[Scalart & Benamar 95] Scalart, P. and Benamar, A.: "On the influence of front-end processing schemes on the GSM coder behaviour in the context of hands-free radiotelephony", *4-th Int. Workshop on Acoustic Echo and Noise Control*, Trondheim, Norway, Norwegian Inst. Technol., pp. 83-86, June 1995

[Schütze 89] Schütze, H.: "Vergleichende Untersuchung von LMS-Algorithmus und Fast-Kalman-Algorithmus zur adaptiven Kompensation akustischer Echos", *Tech. Ber. FI-DBP TELEKOM*, FI 444 TB 16, pp. 1-59, July 1989

[Schütze & Ren 92] Schütze, H. and Ren, Z.: "Numerical characteristics of fast recursive least squares transversal adaptation algorithms - A comparative study", *Signal Processing*, vol. 27, no. 3, pp. 317-332, June 1992

[Schneider 94] Schneider, K. J., Editor: *Bautabellen für Ingenieure mit europäischen und nationalen Vorschriften*, 11-th edition, Düsseldorf: Werner-Verlag GmbH, 1994

[Schultheiss 88] Schultheiss, U.: *Über die Adaption eines Kompensators für akustische Echos*, Dissertation, Technische Hochschule Darmstadt, in *Fortschrittsberichte, VDI Reihe 10: Informatik und Kommunikationstechnik*, Nr. 90, Düsseldorf: VDI Verlag, 1988

[Siemens AG 97] Siemens AG Berlin und München, Editor: *Technische Tabellen, Größen, Formeln, Begriffe*, Edition 1998, Erlangen: Publicis MCD Verlag, 1997

[Siemens AG 98] Siemens AG, Semiconductor Group: "Mobile Communications ICs", *http://www.siemens.de/semiconductor/products/ics/33/3301.htm*, Nov. 22, 1998

[Silverman 87] Silverman, H.: "Some analysis of microphone arrays for speech data acquisition", *IEEE Trans. on ASSP*, vol. ASSP-35, no. 12, pp. 1699-1712, Dec. 1987

[Smolka 94] Smolka, P.: "GSM-Funkschnittstelle" in *GSM-Mobilfunk-Übertragungstechnik*, Preibisch, H., Editor, Berlin: Schiele & Schön, 1994, pp. 32-52

[Sondhi & Kellermann 92] Sondhi, M.M. und Kellermann, W.: "Adaptive echo cancellation for speech signals" in *Advances in Speech Signal Processing*, Sondhi, M. M. und Furui, S., Editors, Third edition, New York: Manual Dekker Inc., 1992, pp. 327-356

[Spencer 98] Spencer, N.: "An overview of digital telephony standards", *IEE Colloquium on Design of Digital Cellular Handsets*, London, UK, pp. 1/1-1/4, March 1998

[Srinivasan & Gersho 93] Srinivasan, K. and Gersho, A.: "Voice activity detection for cellular networks", *Proc. of the IEEE Speech Coding Workshop*, pp. 85-86, Oct. 1993

[Stöcker 95] Stöcker, H., Editor: *Taschenbuch mathematischer Formeln und moderner Verfahren*, Third edition, Thun and Frankfurt am Main: Verlag Harri Deutsch, 1995

[Stearns & David 93] Stearns, S.D. and David, R.A.: *Signal Processing Algorithm Using Fortran and C*, Englewood Cliffs, NJ: Prentice Hall, 1993

[Steele 92] Steele, R.: *Mobile Radio Communications*, First Edition, London, England: Pentech Press Limited, 1992

[Tanaka et al. 95a] Tanaka, M., Kaneda, Y., Makino, S. and Kojima, J.: "Fast projection algorithm and its step size control", *Proc. of ICASSP 95*, Detroit, MI, USA, pp. 945-948, May 1995

[Tanaka et al. 95b] Tanaka, M., Kaneda, Y., Makino, S. and Kojima, J.: "A fast projection algorithm for adaptive filtering", *Trans. IEICE Japan*, vol. E78-A, no. 10, pp. 1355-1361, Oct. 1995

[Taylor 94] Taylor, S.: "Silencing the echo", *Commun. Int. (UK)*, vol. 21, no. 2, pp. 27-33, Feb. 1994

[Tucker 92] Tucker, R.: "Voice activity detection using a periodicity measure", *IEE Proceedings-I Communications, Speech and Vision*, vol. 139, no. 4, pp. 377-380, Aug. 1992

[Vaidyanathan 93] Vaidyanathan, P. P.: *Multirate Systems and Filter Banks*, Englewood Cliffs, New Jersey: Prentice-Hall, 1993

[Van Compernolle 92] Van Compernolle, D.: "DSP techniques for speech enhancement", *Proc. of ESCA Workshop on Speech Processing in Adverse Conditions*, Cannes, pp. 31-42, Nov. 1992

[Vary 85] Vary, P.: "Noise suppression by spectral magnitude estimation - mechanism and theoretical limits", *Signal Processing*, vol. 8, pp. 387-400, 1985

[Vaseghi 96] Vaseghi, S.V.: *Advanced Signal Processing and Digital Noise Reduction*, New York: John Wiley & Sons Ltd., Leipzig: B.G. Teubner, Verlagsgesellschaft mbH, 1996

[Walker 93] Walker, M.: "Handsfree Speaking - A Step Towards Natural Communication", *Electrical Communications*, no. 2, pp. 181-187, 1993

[Walke 98] Walke, B.: *Mobilfunknetze und ihre Protokolle. Band 1: Grundlagen, GSM, UMTS und andere zellulare Mobilfunknetze*, Stuttgart, Germany: Teubner, 1998

[Waters 91] Waters, G.: "Speech production and perception" in *Speech processing*, Rowden, C., Editor, London: McGraw Hill Book Company, 1991, pp. 1-34

[Weiss et al. 91] Weiss, C., Berlin, F. and Geller, D.: "Advanced voice communication with acoustic echo cancelling and enhanced sound quality", *Int. Conf. on Communications 91*, vol. 1, pp. 478-483, Denver, USA, June 1991

[Widrow & Stearns 85] Widrow, B. and Stearns, S.D.: *Adaptive Signal Processing*, Englewood Cliffs, New Jersey: Prentice-Hall, 1985

[Widrow et al. 75] Widrow, B., Glover jr., J.R., McCool, J.M., Kaunitz, J., Williams, C.S., Hearn, R.H., Zeidler, J.R., Dong jr., E. and Goodlin, R.C.: "Adaptive noise cancelling: principles and applications", *Proc. of IEEE*, vol. 63, no. 12, pp. 1692-1716, Dec. 1975

[Xydeas et al. 88] Xydeas, C., Baghbadrani, D. and Erwood, A.: "Speech processing in mobile radio communications", *IEEE Colloquium on Digitized Speech Communication via Mobile Radio*, London, UK, no. 139, pp. 3/1-12, Dec. 1988

[Yang 93] Yang, J.: "Frequency domain noise suppression approaches in mobile telephones systems", *Proc. of ICASSP 93*, Minneapolis, USA, vol. 2, pp. 363-366, April 1993

[Yasukawa 92] Yasukawa, H.: "Acoustic echo canceller with sub-band noise cancelling", *Electronic Letters*, vo. 28, no. 15, pp. 1403-1404, July 1992

[Ye & Wu 91] Ye, H. and Wu, B.X.: "A new double-talk detection algorithm based on the orthogonality theorem", *IEEE Trans. Comm.*, vol. 39, no. 11, pp. 1542-1545, Nov. 1991

[v.Zitzewitz 89] von Zitzewitz, A.: *Annäherung an das ideale Freisprechtelephon mittels Nachbildung der Übertragungsstrecke Lautsprecher-Raum-Mikrophon*, Doctoral Dissertation, Ruhr-Universität Bochum, 1989