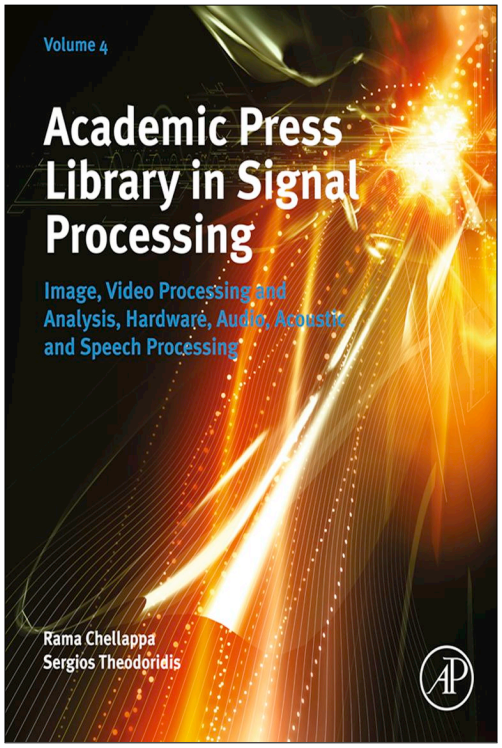
Provided for non-commercial research and educational use only.
Not for reproduction, distribution or commercial use.

This chapter was originally published in the book *Academic Press Library in Signal Processing*. The copy attached is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research, and educational use. This includes without limitation use in instruction at your institution, distribution to specific colleagues, and providing a copy to your institution's administrator.



All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:
http://www.elsevier.com/locate/permissionusematerial

From Gerald Enzner et al., Acoustic Echo Control. In: Rama Chellappa and Sergios Theodoridis, editors, Academic Press Library in Signal Processing. Vol 4, Image, Video Processing and Analysis, Hardware, Audio, Acoustic and Speech Processing, Chennai: Academic Press, 2014, p. 807-877.
ISBN: 978-0-12-396501-1
© Copyright 2014 Elsevier Ltd
Academic Press.

Author's personal copy

CHAPTER

Acoustic Echo Control

30

Gerald Enzner[*], Herbert Buchner[†], Alexis Favrot[‡] and Fabian Kuech[§]

[*] *Ruhr-Universität Bochum, Universitätsstraße 150, D-44780 Bochum, Germany*
[†] *Machine Learning Group, Technische Universität Berlin, Franklinstraße 28/29, D-10587 Berlin, Germany*
[‡] *Illusonic LLC, Chemin du Trabandan 28A, CH-1006 Lausanne, Switzerland*
[§] *Fraunhofer Institut für Integrierte Schaltungen, Am Wolfsmantel 33, D-91058 Erlangen, Germany*

Nomenclature

List of mathematical symbols

The following general conventions are used in this chapter: Matrices and vectors are written boldface, e.g., $\mathbf{x}(n)$, matrices often uppercase and vectors lowercase, while scalar sequences are written non-bold, e.g., $x(n)$. Quantities in the frequency-domain are written in uppercase Latin, e.g., $X(k, m)$, or boldface in the DFT-domain, e.g., $\mathbf{S}(m)$. Estimated quantities are labeled with a hat, e.g., $\hat{s}(n)$. Some special quantities are denoted by Greek symbols:

$d(n)$	acoustic echo signal
$e(n)$	error signal after acoustic echo cancellation
f_s	sampling frequency
$G(m, \nu)$	echo path transfer function (delay compensated!)
h_n	acoustic echo path coefficients at time-lag n
$\mathbf{I}$	identity matrix
$\mathbf{k}(n)$	time-varying Kalman gain vector
m	block time index, $m \in \mathbb{Z}$
n	discrete sampling time index, $n \in \mathbb{Z}$
N	number of FIR filter coefficients
$\mathbf{R}_s(n)$	time-varying covariance matrix of near-end speech s
$s(n)$	near-end speech signal (including environmental noise)
$x(n)$	received signal from the far speaker
$y(n)$	microphone signal
μ	stepsize factor of LMS-type algorithms
ν	discrete frequency index, $\nu = 0, 1, \ldots, M-1$
$\sigma_s^2(n)$	time-varying power, i.e., variance, of near-end speech $s(n)$
$(\cdot)^H$	transposition and complex conjugation of the argument
$(\cdot)^T$	transposition of the argument

Academic Press Library in Signal Processing. http://dx.doi.org/10.1016/B978-0-12-396501-1.00030-3
© 2014 Elsevier Ltd. All rights reserved.

Author's personal copy

$\mathcal{E}\{\cdot\}$	statistical expectation operator
ln, log	natural and base-10 logarithm

List of Abbreviations

A/D	Analog-to-Digital
AEC	Acoustic Echo Control
AES	Acoustic Echo Suppression
AGC	Adaptive (or Automatic) Gain Control
APA	Affine Projection Algorithm
DCR	Diagonal Coordinate Representation
DFT	Discrete Fourier Transform
EEF	Echo Estimation Filter
EOS	Equivalent Orthogonal Structure
ESF	Echo Suppression Filter
ERLE	Echo Return Loss Enhancement
FDAF	Frequency-Domain Adaptive Filter
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
IDFT	Inverse Discrete Fourier Transform
IFFT	Inverse Fast Fourier Transform
IIR	Infinite Impulse Response
IP	Internet Protocol
ITU	International Telecommunication Union
LMS	Least Mean-Square Algorithm
MCAEC	Multichannel Acoustic Echo Cancellation
MMSE	Minimum Mean-Square Error
NLMS	Normalized Least Mean-Square Algorithm
PSD	Power Spectral Density
RLS	Recursive Least-Squares Algorithm
SER	Signal-to-Echo Ratio
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
VAD	Voice Activity Detector

4.30.1 Introduction

Acoustic echo control (AEC) refers to signal processing technology used in communication systems with *full-duplex* hands-free acoustic man-machine interfaces. AEC technology is required to combat the undesired acoustic coupling between sound reproduction and acquisition in a system. In this introductory section, we first clarify the problem statement, before we describe typical elements of the

Author's personal copy

signal processing chain in sound front-ends of hands-free communication systems. We then refer to most relevant applications and quality measures for AEC technology.

4.30.1.1 Problem statement and early developments

In the signal model of a hands-free voice communication system in Figure 30.1, the task of the AEC unit is to reproduce the desired near-end speech $s(n)$, at discrete time n, in the output signal $\hat{s}(n)$ in sending direction of the system, while suppressing the undesired echo $d(n)$ of the far-end speech $x(n)$. The echo signal due to the echo path h_n typically consists of the direct sound between loudspeaker and microphone as well as multiple reflections of walls in the enclosure. The larger the roundtrip delay of the far-end speech in an end-to-end communication, the more echo attenuation is required [1].

The black-box AEC in Figure 30.1 represents different types of signal processing that have been pursued to achieve the separation of the echo from the near-end speech. *Voice controlled switching* has been developed in the 1960s and is still used in many products to suppress the acoustic echo of the far speaker. In an example of this technique, the hands-free microphone signal $y(n)$ is strongly attenuated whenever a received signal $x(n)$ from the far speaker side is observed. Alternatively, we could attenuate the received signal $x(n)$ before reproduction by the loudspeaker if near-end speech is detected and has priority. In any case, the loop gain from $x(n)$ to $\hat{s}(n)$ should provide the required echo attenuation. Voice controlled switching can be implemented easily, analog or digital, but the fundamental problem is that switching effectively leads to an unacceptable half-duplex connection between both ends of the communication system. Especially the background noise transmission would suffer from on-off modulations. Therefore, voice controlled switches (or other gain functions in the system) are nowadays implemented in conjunction with *comfort noise injection* [2].

Many of the current developments in AEC can be traced back to the fundamental idea of creating a replica of the echo path impulse response h_n in Figure 30.1 using FIR filter structures. The FIR filter is essentially placed in parallel to the echo path and stimulated by the loudspeaker signal.

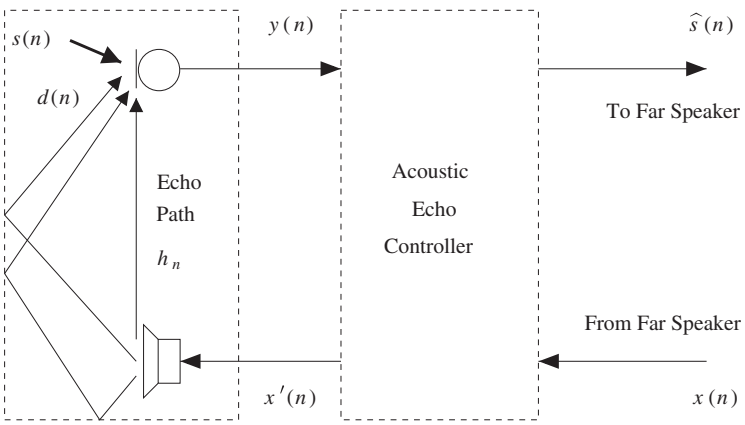


FIGURE 30.1

Acoustic front-end of hands-free voice communication systems.

Author's personal copy

The resulting echo replica at the filter output can be subtracted from the microphone signal to achieve echo cancellation and said FIR filter is thus termed *echo canceler*. This principle was originally conceived in the 1960s to solve the problem of *line echoes* due to non-ideal hybrids in long distance telephone networks [3,4]. Using this technology for canceling *acoustic* echoes, however, turns out to be a quite challenging task, since the duration of the impulse response of an acoustical echo path is usually many times longer than impulse responses of line echoes. Depending on the acoustic environment and the sampling frequency, acoustic echo cancellation then requires FIR filters with a very large number of coefficients. In the small interior of the car environment and for the low sampling frequency of 8 kHz, i.e., for telephone quality, it already amounts to hundreds of coefficients, while in the office environment we may even encounter thousands of taps depending on the reverberation time [4]. It has been shown that IIR filters will not achieve better echo path modeling with lesser coefficients [5].

4.30.1.2 Typical building blocks of current systems

Virtually all existing systems are based on the early ideas of echo attenuation and echo cancellation. In practice, it turns out that hybrid solutions often provide the best performance. The mainstream in system development for acoustic echo control is wrapped up in Figure 30.2, which basically represents a zoom into the black-box AEC of Figure 30.1. If the echo canceler would exactly match the echo path, then the echo signal could be eliminated completely from the microphone signal. However, echo paths are in most of the cases *a priori* unknown and time-varying systems (e.g., due to moving persons in the vicinity of loudspeaker and microphone) and for this reason adaptive filters and adaptive filter algorithms have to be considered. The developments for the *adaptive echo canceler* are reported in great detail, e.g., in [6–9] and references therein.

Due to the related computational load and numerical requirements, the field of acoustic echo cancellation eventually has been understood as an application of very-high-order adaptive filters [6]. In this way, it has also been an important driving force in the development of new and powerful adaptive filtering algorithms. Among the desirable properties of adaptive echo cancellation filters, we have low computational complexity, low memory requirements, fast convergence, and excellent tracking ability. Due to this multitude of requirements which have to be satisfied simultaneously, the first successful implementations and commercial products for the acoustic echo case were available not until the 1980s [4]. Again considering Figure 30.2, the most frequently used adaptive filter algorithms in the echo cancellation context are the gradient-based NLMS, APA, RLS, and FDAF type of algorithms. Their properties are thoroughly described for example in [10] and briefly summarized here:

- The *normalized least-mean-square* (NLMS) algorithm is by far the most popular technique for adaptive identification of the acoustic echo path [2]. The popularity of the NLMS algorithm is due to its simplicity, low complexity, and robust (i.e., model-independent) performance. In the case of correlated (i.e., non-white) echo path input signals $x(n)$, such as speech, the convergence rate of the NLMS algorithm is, however, too slow to track the echo path impulse response of time-varying acoustic environments. This problem can be resolved partly by applying decorrelation filters to the input $x(n)$ and the echo cancellation error $e(n)$ before feeding them into the NLMS algorithm [11,12].
- The *affine projection algorithm* (APA) introduced in [13] utilizes several previous input vectors to determine the input signal correlation. The APA can be seen as a generalization of the NLMS

Author's personal copy

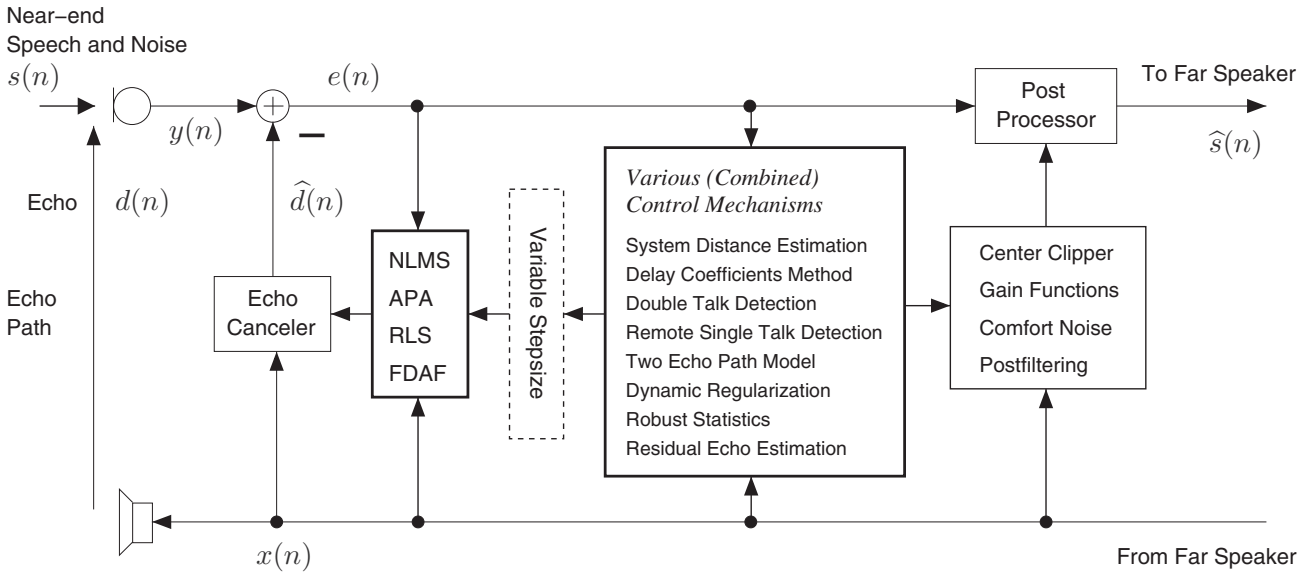


FIGURE 30.2

Elements of a hands-free communication system.

Author's personal copy

algorithm and converges rapidly for autoregressive input signals [14]. A computationally efficient realization has been presented in [15].

- The fastest convergence is achieved by the *recursive least-squares* (RLS) algorithm. For decorrelation purposes, the RLS algorithm utilizes the inverse of the autocorrelation matrix of the input signal, the numerical computation of which is demanding and needs to be handled with care [16]. Computationally efficient implementations have been developed [17,18], but still the RLS algorithm is often considered as too cumbersome for practical applications.
- Subband solutions and *frequency-domain adaptive filters* (FDAF) have been developed as interesting alternatives with good convergence properties, low computational complexity, and often favorable numerical properties [19–21]. The drawback inherent to most of these techniques is the signal delay that is caused by analysis and synthesis filtering or windowing. The signal delay related to the FDAF can be reduced either by using the partitioned-block frequency-domain adaptive filter (PBFDAF)/multi-delay filter (MDF) [22–24], or the specific soft-partitioned frequency-domain adaptive filter (SPFDAF) [25], or simply by choosing large frame overlap in the FDAF overlap-save architecture, e.g., [26–28], the latter at the expense of larger computational load. In all these variants of the FDAF, the algorithmic signal delay and the actual filter length can be adjusted independently. Delayless subband adaptive filters, which essentially perform the adaptation process in the frequency-domain and the actual filtering in the time-domain, are described, e.g., in [29,30].
- Another class of algorithms, with increasing popularity, exploits the sparseness of impulse responses in both network and acoustic echo cancellation applications. In particular, the sparseness is taken into account by updating the filter coefficients independently of each other with different adaptation time constants—in *proportion* to the magnitude of the already estimated filter coefficients. Implementation of this proportionate mechanism obviously requires another feedback loop in the algorithm, besides the already utilized output error signal. Larger coefficients are then adjusted with faster convergence, while smaller coefficients converge slower, and consequently the overall speed of convergence increases. One of the first proportionate algorithms is found in [31]. Since that time, variants and improvements of the basic proportionate NLMS (PNLMS) algorithm have been presented, e.g., [32,33]. All this exploitation of sparseness in echo cancellation should be viewed in the context of seminal work on sparse regression [34].

Adaptive algorithms as mentioned here have two (related) fundamental problems in common: tracking ability and robustness. It turns out difficult to let the echo canceler coefficients follow the true time-varying echo path of realistic acoustic environments. On the one hand, the adaptation must be fast enough to track the time-varying system, while on the other hand, the adaptation must be robust against interfering near-end speech (so-called double-talk situation) and background noise. Both requirements are contradicting and herein lies the key problem of acoustic echo control. As a result, sophisticated control mechanisms were proposed to support the fast and robust adaptation of echo canceler coefficients [2,35]. A perfect solution is, however, not available due to the nature of a statistical optimization problem. In an attempt to satisfy both requirements, many systems utilize *time-varying adaptive stepsize control* to accelerate the adaptation in the absence of near-end speech and noise and to slow down (or even halt) otherwise. Since there are as many publications on stepsize control as on adaptive filters, we can refer to only some contributions.

Author's personal copy

An optimum stepsize for the NLMS algorithm has been determined by quite some authors [2,10,36–39], where the optimization is always based on the idea of minimum mean-square system distance between echo canceler and echo path. Optimum stepsizes for the FDAF and the PBFDAF have been derived in [40]. Unfortunately, all these stepsizes cannot be implemented directly as the (frequency-dependent) time-varying system distance between echo canceler and echo path is also required as an input parameter, but difficult to measure out of the lab. Thus, suboptimal control mechanisms have been developed to approximate the optimum stepsize. Some methods have been designed to explicitly estimate the actual *system distance*, whereas other methods aim to facilitate system distance estimation or to control the adaptation directly:

- The popular *delay coefficients method* computes the system distance from the leading coefficients of the adaptive filter, provided that the echo path has natural or artifical zeros at the corresponding impulse response lags [39]. It is important to note that the delay coefficients method alone is not able to deliver a reliable estimate of the system distance. An additional detector for echo path changes is required to avoid stalling (freezing) of the adaptation [2,10,36].
- *Double talk detectors* (DTD) can be utilized to directly halt the adaptation in the presence of near-end speech at the microphone [35] or, alternatively, to facilitate system distance estimation [2]. DTD can be based on cross-correlation measures [41] or on the simple comparison of signal powers or magnitudes [42]. *Remote single talk detection (voice activity detection)* as described in [2] is closely related to DTD.
- The *two echo path model* [43] can be used in different ways to control the adaptation. Basically, the approach models a fast and a slowly changing echo path by a background and a foreground adaptive filter, respectively. An evaluation and comparison of their respective echo cancellation errors then provides means for adaptive stepsize control in the foreground adaptive filter. This technique finds widespread application in cases where its computational complexity is tolerable.
- *Dynamic regularization* controls gradient adaptive filters by means of a time-varying additive quantity in the denominator of the gradient [10,44,45]. It was shown in [46] for the NLMS algorithm that optimum regularization is equivalent to optimum stepsize control.
- The term *residual echo power estimation* [47–50] basically refers to the same thing as *system distance estimation*, since the respective quantities are simply related by the available echo path input signal power. System distance estimation is usually preferable, as it separates system properties from signal properties.

In recent years, in addition to the above control mechanisms, adaptation algorithms have been developed which are inherently more robust to double talk. Instead of the least-squares approach as in NLMS, APA, RLS (which corresponds to Gaussian noise assumption) they are based on more advanced stochastic signal models. Basically, there are three fundamental stochastic signal properties that may be taken into account in these models, and hence, be exploited by the adaptation algorithm: *nonstationarity*, *nonwhiteness*, and *nongaussianity*.

The simplest way of exploiting nongaussianity—which is equivalent to taking into account higher-order statistics (HOS)—is to apply a suitable error nonlinearity within the optimization criterion. In the context of echo cancellation the error nonlinearity was already mentioned in the early paper [3] in order to cope with *outliers* caused by misdetections of the double-talk detector. The outliers generally exhibit a supergaussian probability density. Later, the concept of *robust statistics* [51], known from

Author's personal copy

the statistics literature to systematically address the need of outlier robustness, has been identified as a powerful and systematic statistical framework for the control of adaptive filters [35,52,53]. It turns out that this concept can in fact be seen as a systematic and statistically motivated refinement of the use of an error nonlinearity together with a certain adaptive scaling factor.

Another important and even more general class of HOS-based statistical methods is independent component analysis (ICA) [54] which can be interpreted from an information-theoretic point of view, and was originally developed for the problem of blind source separation. Considering now the *acoustic echo control problem from this information-theoretic point of view*, an alternative criterion for acoustic echo control would be to minimize the mutual information between the transmission signal $\hat{s}(n)$ according to Figure 30.1 and the loudspeaker signal $x'(n)$. In other words, the transmission signal $\hat{s}(n)$ should be made statistically independent from the echo contribution. Hence, from this point of view, AEC can be considered as a (supervised) signal separation problem with the accessible loudspeaker signal as side information. Indeed, using a generic broadband formulation of ICA for convolutive mixtures, a conceptually simple, yet fundamental relation between AEC and blind adaptive filtering algorithms was established in this way in [55,56].

In addition to the nongaussianity, the *nonstationarity* and the *nonwhiteness* of the signals can be expoited by the adaptation algorithm. Indeed, all the above-mentioned double-talk handling approaches are based directly on the nonstationarity property of the near-end speech signal $s(n)$, and in some cases on the nonwhiteness of the involved signals. Both the nonwhiteness and nonstationarity are captured by (time-varying) correlation matrices in multivariate second-order statistics.

In [55,57] all three fundamental approaches, i.e., exploiting nongaussianity, nonwhiteness, nonstationarity, are combined in one generic information-theoretic broadband adaptive filtering framework, which we call TRINICON ("TRIple-N Independent component analysis for CONvolutive mixtures"). An interesting finding from this top-down approach is that the generic framework contains not only all well-known adaptation algorithms mentioned above (e.g., NLMS, APA, RLS) but it also inherently includes the various known adaptation controls mentioned above, including robust statistics [55,56]. This has led to various new insights into the adaptation mechanisms and we believe that it continues to be a valuable source of synergy effects for the development of new and improved adaptation algorithms.

In addition to the modeling and exploitation of these *signal* properties, later parts of this chapter will be concerned also with introducing the somewhat younger treatment of *system* properties, especially system uncertainties. Thereby we will mostly focus on minimum mean-square error (MMSE) optimization based on Gaussian statistics, i.e., the special case of second-order statistics. A joint consideration of both signal and system properties has not been achieved yet and may thus represent a future endeavor.

Despite the availability of fast and robust adaptive echo cancelers, residual echo always remains after the echo canceler and it is widely accepted that an echo canceler alone will not be able to deliver sufficient echo attenuation in all situations. Adaptive echo cancellation combined with voice controlled switching is thus implemented in real systems to improve the echo attenuation, but the distortion of the desired signal $s(n)$ due to switching can still be unacceptable. Another major branch of developments in the field of acoustic echo control is thus devoted to more sophisticated *post-processing* for residual echo suppression in the sending path of the communication system, as shown by Figure 30.2. In stand-alone form, i.e., without an echo canceler in the system, we refer to *echo suppression* technology. The latter is most applicable if the many requirements to the adaptive echo canceler (such as low complexity,

Author's personal copy

fast tracking ability, clock synchronization etc.) cannot be satisfied in an application at hand. Further motivation for stand-alone *echo suppression* will be provided in Section 4.30.3.

Frequency-selective post-processing, in conjunction with echo cancellation, is also termed *postfiltering* for residual echo suppression [58–61]. In principle, the operation of a residual echo suppression postfilter is very similar to that of a noise suppression filter and it was reported that both functionalities can be combined efficiently [58,62]. The key for a good postfilter is the availability of the power spectral density (PSD) of the residual echo signal. As the residual echo is not a measurable signal, the estimation of the residual echo PSD from the available signals has to be designed carefully [48,49].

Some publications have shown a tight relationship between the optimum statistical adaptation of echo canceler and postfilter coefficients [48,63,64]. Ideally, both filters employ the residual echo PSD as a control parameter. On the one hand, the residual echo PSD governs the adaptive stepsize of LMS- and FDAF-type adaptive echo cancellation filters and, on the other hand, it is required in the spectral weight calculation for adaptive postfiltering. Exploitation of this relationship leads to intelligent interaction of both filters with shared responsibility for acoustic echo control. Interestingly, joint control of echo canceler and postfilter can be realized simpler than their individual control, if the aforementioned synergy is taken into account. This has been demonstrated, e.g., by the compact and contained algorithm in [27], which provides the required echo attenuation and also preserves the full-duplex ability of the system [65].

Regarding the post-processor in general, there are further options how to adjust a fixed or adaptive, linear or nonlinear, scalar or frequency-dependent echo attenuation. Post-processing techniques are, however, not so well documented in the literature as adaptive filters for the echo canceler. Byproducts of the control mechanisms for the echo canceler are sometimes used to adjust the post-processor. More hints on post-processors can be found for example in [2,35,66, Chapter 7].

Yet more algorithms and variants of the aforementioned algorithms have been proposed in the literature, so that our presentation can never be complete. Further references are provided in the course of this chapter where applicable in the respective context. For complete bibliography, including the history of acoustic echo control, we recommend further reading in [2,8,9,35].

4.30.1.3 Applications of acoustic echo control

The most important business for acoustic echo control technology is created by hands-free telephones and speech dialog systems with full-duplex ability. In this chapter, we are mainly concerned with algorithms for hands-free telephony in different environments. Recently, the hands-free telephone market has been growing due to the advent of modern telecommunication systems, such as

- mobile phones and smart phones with integrated or external hands-free loudspeaker unit,
- car hands-free telephones (integrated or based on mobile phone),
- desktop teleconferencing equipment (e.g., dedicated hardware or PC based solutions for Voice-Over-IP),
- and audio-visual environments for tele-presence (e.g., HP Halo, Cisco Telepresence, or Tandberg T3).

The disturbing effect of the acoustic echo and the requirement for acoustic echo control essentially arise from large transmission delays in modern communication. The echo signal delay (which is twice the transmission delay) ranges from about 200 ms in mobile radio up to seconds in some

Author's personal copy

IP connections. The large transmission delay of IP connections necessitates acoustic echo control even in handset or headset modes of communication. Low signal-to-noise ratio at the microphone, such as in car hands-free telephony and mobile phones, further causes uncertainties regarding the precise detection and separation of the disturbing echo from the desired signal.

In many of the aforementioned configurations, the hands-free voice interface is supposed to provide for improved user-friendliness and simplicity of the communication. In the area of car telephony, the motivation for hands-free solutions is not only given by the additional user convenience. Here, it is rather the case that legal aspects in the form of safety regulations are the driving force for the use of hands-free systems.

Speech dialog systems refer to applications which make use of automatic speech recognition (ASR) units. Originally, speech recognition was mostly relevant for industrial applications, automatic answering services over the phone, and smart traffic products (navigation). More recently, the availability of modestly priced processing power has also made the ubiquitous ASR possible in smart home applications, i.e., in the consumer market. In speech dialog systems with full-duplex ability, acoustic echo control based on the known reproduction signal is required as a preprocessing unit for ASR to avoid confusion between user commands and simultaneous loudspeaker output. This line of research is of high practical importance for various kinds of next-generation multimedia terminals, such as advanced TV sets, future multimedia workstations, systems for interactive retrieval of multimedia data, computer games, navigation, and other voice controlled systems with the ambition of continuous recognition and reproduction.

4.30.1.4 Quality measures

According to product advertisement in the hands-free telephone market, the speech quality of the respective solution is always excellent, i.e., the sound is *loud* and *clear* and the system has *full-duplex ability* (simultaneous transmission in both directions). Of course, such statements are derived from well-known customer needs and, therefore, they give us a first insight into the quality aspects of hands-free telephones. What the customer in fact expects is a quality that cannot be distinguished from the hand-held telephone. Quality impairments are usually not acceptable and can immediately lead to a reduced acceptance of the product in the market. This has been recognized as a serious problem, e.g., for premium car manufacturers if their pre-installed hands-free telephones do not fully comply with quality expectations.

The reality of hands-free telephones shows that in quite some cases the speech quality is not sufficient yet. Typical complaints of users refer to a "thin and metallic (or reverberant) sound of the voice", "annoying half-duplex functionality", "insufficient loudness", or "low intelligibility due to noise". The same complaints were also confirmed by the results of independent test labs for car hands-free telephones, e.g., [67,68]. The lack of quality then explains an insufficient consumer acceptance of hands-free products, as for example in the automotive area—despite the hand-held phone ban enforced by legislation.

Objective prediction and evaluation of the subjective quality of modern hands-free communication systems is a very complex issue and is still under investigation, e.g., [69]. The difficulty is due to a variety of nonlinear and time-variant signal processing in the hands-free terminals and in the network (e.g., dynamic level control, echo cancellation, residual echo suppression, noise reduction, comfort noise injection, coding, jitter buffering, and error concealment). Thus, the description of the perceived

Author's personal copy

speech quality by a single value or quality index seems to be out of sight. What can be said is that the subjective quality of hands-free telephones depends at least on the following quality parameters:

- loudness and sound quality, i.e., listening speech quality,
- acoustic echo attenuation, i.e., talking speech quality,
- double talk speech quality, i.e., duplex ability,
- and naturalness of (residual) background noise transmission.

At this point, it should be noted that the *simultaneous optimization* of these quality parameters is indeed required to achieve sufficient performance. This is the requirement which constitutes the actual difficulty in the field of acoustic echo control. Relaxation of just one or two quality aspects stands for undue simplification in the design of hands-free telephones. For instance, by significantly reducing the reproduction loudness of the device, we can easily avoid echo, but obviously it does not help the communication.

Instrumental speech quality measures, which have been described, e.g., in [70,71], and newer techniques which have been developed for the analysis of coded and transmitted speech, e.g., [72–74], are not directly applicable to evaluate the quality of acoustic echo control and speech enhancement systems in general. The reason is that the specific signal modifications and distortions which are introduced by acoustic echo and noise reduction techniques are not explicitly modeled in these approaches.

A variety of subjective and objective quality parameters, as well as testing methods for systems which rely on acoustic echo control, are described in the following ITU-T recommendations: [1,75–82]. A comprehensive and promising specification of objective test procedures for car hands-free telephones has been enforced by the association of German car manufacturers (VDA) [83,84]. The VDA specification is based on previously mentioned ITU-T recommendations, but in addition to the standard quality parameters, the VDA defines a more detailed analysis of double talk situations and background noise transmission.

From the algorithm developer's viewpoint in the field of acoustic echo control, a convenient way of assessing the echo quality of a hands-free system is the calculation of the echo return loss enhancement (ERLE) as an indicator function for the echo attenuation achieved by the system. Based on the signals at sampling time index n in the block diagram in Figure 30.2, the ERLE can be defined via statistical expectation $\mathcal{E}$ as follows:

$$\mathrm{ERLE} = \frac{\mathcal{E}\{d^2(n)\}}{\mathcal{E}\{(d(n) - \hat{d}(n))^2\}}. \tag{30.1}$$

This formula can be applied under lab conditions when the echo signal $d(n)$ and thus the residual echo $d(n) - \hat{d}(n)$ after echo cancellation are available explicitly. Otherwise, the echo signal $d(n)$ can be replaced by the noisy microphone signal $y(n)$ and conclusions regarding the echo attenuation can be drawn as long as the level of near-end speech and noise is low. After the post-processor in Figure 30.2, a more suitable measure is to evaluate the resulting near-end speech quality in form of a signal-to-echo ratio (SER) according to the following definition:

$$\mathrm{SER} = \frac{\mathcal{E}\{s^2(n)\}}{\mathcal{E}\{(s(n) - \hat{s}(n))^2\}}. \tag{30.2}$$

Author's personal copy

Besides the estimated speech $\hat{s}(n)$, the isolated near-end speech signal $s(n)$ needs to be available for SER calculation. Both measures, ERLE and SER, can be represented as time-varying functions when the statistical expectation is resolved by short-time averaging.

As shown by previous sections, most AEC systems internally rely on an adaptive estimate $\hat{h}_n$ of the acoustic echo-path impulse response h_n in Figure 30.1 to achieve echo cancellation. In that respect, a normalized *echo path misalignment* measure, often termed *system distance* or *coefficient error norm*, according to

$$D = \frac{\sum_n \|h_n - \hat{h}_n\|^2}{\sum_n \|h_n\|^2}, \tag{30.3}$$

can be very useful for the developer to look inside the system and to prove operation of the adaptive echo path identification algorithm, e.g., NLMS. Under specific circumstances, i.e., in particular with white noise excitation $x(n)$, the echo path misalignment and the echo return loss enhancement turn out to be the inverse of each other [28]. This can be easily recognized by looking at $d(n)$ and $\hat{d}(n)$ as a convolution of h_n and $\hat{h}_n$, respectively, with the same input signal $x(n)$. However, for correlated input signals, such as speech, or multichannel systems with coherent input signals on different input channels, cf. Section 4.30.4, the equivalence of both measures is lost.

An open issue regarding the currently available instrumental and automatic test procedures is, however, that the correlation between predicted speech quality and perceived quality is not always guaranteed. In order to surpass this issue, ongoing developments take the following strategies into account to increase the significance of existing quality measures:

- analysis of the acoustic echo control performance in the case of time-varying echo paths,
- specification of the required echo attenuation in the presence of noise,
- and the measurement of speech quality parameters during the simultaneous presence of speech, noise, and echo.

This brief survey on speech quality of hands-free telephones has shown that research in this difficult area cannot be considered as finalized. Due to its multi-dimensional nature, the objective prediction of speech quality is probably yet more challenging than the speech enhancement problem itself. A more comprehensive overview about the current state-of-the-art in the field of advanced speech quality testing for modern (hands-free) telecommunication systems can be found in [85]. A recent update and some specific directions for future work are presented in [86,87].

4.30.1.5 Outline of this chapter

After this itemized introduction to history and mainstream technology in acoustic echo control, we shall now proceed and deepen the understanding in formal terms. In the core sections of this chapter, we will partly revisit the established technology and add more specific references, but mainly we will trace the more recent trends due to research beyond 2000.

In Section 4.30.2, we first pick up the single-channel linear AEC problem and present a fresh view in terms of an *uncertainty model* of the acoustic environment and a suitable adaptive algorithm development in a recursive Bayesian estimation style. In this uncertainty framework, particular objectives are given

Author's personal copy

by the rigorous justification of the *adaptive echo cancellation and post-processing* hybrid through joint optimization, and by the unification of adaptive filters and adaptive filter control in the form of statistical *Kalman filtering* techniques.

In Section 4.30.3, the uncertainty model about the echo path is understood harsh enough to drop the echo cancellation concept and stick to *echo suppression* alone. This strategy is justified, for example, when the uncertainty is due to data dropouts in the audio stream or clock desynchronization between sound reproduction and acquisition. For the echo suppression, we then present a dedicated class of algorithms which do not rely on complex echo path impulse identification, but rather on the estimation of a smoothed magnitude-squared echo power transfer function.

In Section 4.30.4, the treatment of acoustic echo cancellation (i.e., no echo suppression) is generalized to the case with multiple reproduction channels. The additional problems of *multichannel acoustic echo cancellation* are fundamentally different from those of traditional single-channel echo cancellation. Not only the computational requirements are higher, but also the identification of more unknowns is naturally more difficult, especially when echo path impulse responses with mutually correlated input signals have to be tackled. As a field of its own, we provide major references to multichannel echo cancellation and specifically highlight the art of decorrelation of the inputs of the acoustic channels. For generality, the treatment will be based on the TRINICON framework which includes inherent adaptation control as mentioned above.

Unfortunately, the linear echo path model (single- or multichannel) does often not capture the reality of today's telecommunication devices which may include cheap audio hardware introducing non-negligible nonlinear distortion into the signal played back by the loudspeaker. Informal listening tests indicate that the human listener tolerates nonlinear distortion up to a level at which it can no longer be modeled and cancelled by linear AEC with sufficient accuracy. As a result, unacceptable nonlinearly distorted acoustic echo would remain in the signal chain. Hence, in Section 4.30.5, we treat the extension to *nonlinear acoustic echo cancellation* in the form of a survey of the current state-of-the-art in this field.

While these core sections describe the fundamental approaches in the respective fields, we finally move on to a more application-oriented presentation of AEC technology and results in Section 4.30.6. In this *application corner*, we are then dealing with different acoustic environments, such as in car hands-free telephony, desktop teleconferencing, living room, and mobile phones. After this, we draw our conclusions from this chapter.

4.30.2 Echo cancellation and postfiltering

Echo cancellation might be considered as the ideal solution to acoustic echo control, since the acoustic echo could be removed without harm to the desired near-end speech. However, it depends on the working assumption that the echo path impulse response is determined with sufficient accuracy. In this section, we develop a generalized view which will explicitly take uncertainty about the acoustic echo path into account. We shall see the consequences of the uncertainty model of the echo path on the optimum signal processing for acoustic echo control, where the optimization of the system will explicitly target the estimation of the desired near-end speech.

Author's personal copy

At first we formulate the uncertainty model of the echo path by means of a multivariate random variable with statistical mean and covariance. The covariance of the random variable basically describes the uncertainty about the true echo path coefficients. From here, we derive the linear minimum mean-square error (MMSE) estimator for the near-end speech components in the microphone signal. The resulting estimator consists of a subtractive echo canceler which duplicates the systematic part of the echo path (i.e., the echo path expectation) and a statistical postfilter for residual echo suppression due to the echo path uncertainty. This result proves, by the presence of uncertainty alone, the coexistence of echo canceler and postfilter for otherwise linear echo path models and unlimited number of echo canceler coefficients. Echo cancellation with postfiltering for residual echo suppression had previously found its justification merely by the presence of nonlinearities or undermodeled impulse response tails of the echo path.

Then we develop a Bayesian adaptive algorithm for joint mean and covariance estimation of the echo path. The derivation is based on an uncertainty model which represents the typical variability of the echo path in the form of a stochastic Markov model. Since this time-varying echo path is observed in the presence of independent near-end speech at the microphone, the Bayesian adaptive algorithm turns out to be a Kalman filter.

Throughout this section, it remains a contrast to other literature that the echo path is modeled as a random process, whereas the known echo path input is treated as a deterministic signal. Nevertheless, the classical Wiener solution for subtractive echo cancellation is included as a special case.

4.30.2.1 Uncertainty model of the linear echo path

In this section, we prepare for the rigorous joint derivation of acoustic echo canceler and postfilter by setting up the uncertainty model of the linear echo path. In the subsequent Section 4.30.2.2, we then derive the MMSE optimum filtering solution for the problem at hand. The core procedures in this and the next section had been outlined in [27,88]. Here, we shall take the opportunity to translate these procedures into unified vector notation in order to smoothly establish the relationship with vector-oriented adaptive filter algorithms for echo path mean and covariance identification in Section 4.30.2.5.

Getting back to the system in Figure 30.1, let us assume that the impulse response h_n entirely models the electroacoustic coupling between the loudspeaker signal $x'(n)$ and the microphone signal $y(n)$. If sufficiently good transducers are used, the linear echo path is widely accepted as a realistic model for the acoustic environment of hands-free systems. To provide transparent sound at least to the near-end speaker, who typically is the owner of the system, we set the loudspeaker signal equal to the received signal,[1] i.e., $x'(n) = x(n)$. The microphone signal $y(n)$ comprising near-end speech $s(n)$ and

[1] In an even more general approach, we may abolish the simplification $x'(n) = x(n)$ and consider some kind of signal processing also in receiving direction of the system. In this case, the relation between the output $x'(n)$ and the inputs $x(n)$ and $y(n)$ has to be defined in a similar way as for the output $\hat{s}(n)$ later in Eq. (30.13). Furthermore, an appropriate distortion measure then has to be defined in receiving direction and has to be minimized simultaneously with the distortion in sending direction. In this way, the distortion in sending direction might be reduced at the expense of a distortion in receiving direction. For the sake of simplicity, we are not dealing with filters in receiving direction here, but it seems promising to treat this issue in future work.

Author's personal copy

convolutive echo $d(n)$ then reads

$$\begin{aligned} y(n) &= s(n) + d(n) \\ &= s(n) + h_n * x(n) \\ &= s(n) + \sum_{k=0}^{N_1} h_k x(n-k) \\ &= s(n) + \mathbf{h}^T \mathbf{x}(n), \end{aligned} \tag{30.4}$$

where the vector

$$\mathbf{x}(n) = \left[x(n), x(n-1), \ldots, x(n-N_1)\right]^T \tag{30.5}$$

denotes a collection of the most recent echo path input samples at discrete time n and

$$\mathbf{h} = \left[h_0, h_1, \ldots, h_{N_1}\right]^T \tag{30.6}$$

is the finite set of the N_1 corresponding impulse response coefficients.

In the traditional theory of acoustic echo control, the speech signal $s(n)$ and the received signal $x(n)$ are both modeled as independent *random processes*, while the echo path h_n is treated as an unknown *deterministic* parameter. For these model assumptions, the minimization of the mean-square output error $e(n)$ of the echo cancellation filter, cf. Section 4.30.1, leads to the well known Wiener solution, i.e., the echo canceler ideally mimics the true echo path in order to compensate for the acoustic echo in the microphone signal [2,10]. This solution will be absorbed as a special case of our uncertainty framework as shown in Section 4.30.2.3.

In the uncertainty framework, to be considered from now, the speech signal $s(n)$ is not observable alone and is therefore still modeled as a stationary *random process* with zero mean and autocorrelation matrix $\mathbf{R}_s = \mathcal{E}\{\mathbf{s}(n)\mathbf{s}^T(n)\}$ based on the length N_2 vector

$$\mathbf{s}(n) = \left[s(n), s(n-1), \ldots, s(n-N_2)\right]^T. \tag{30.7}$$

However, the echo signal $d(n)$ is now given as the linear convolution of a *measurable* (i.e., *deterministic*) loudspeaker signal $x(n)$ with unknown echo path coefficients h_n. Due to this uncertainty about the acoustic echo path, the coefficient vector $\mathbf{h}$, too, is modeled as an independent *random variable* with some statistical expectation $\hat{\mathbf{h}}$ and covariance $\mathbf{p}$:

$$\hat{\mathbf{h}} = \mathcal{E}\{\mathbf{h}\}, \tag{30.8}$$

$$\mathbf{h}_r = \mathbf{h} - \hat{\mathbf{h}}, \tag{30.9}$$

$$\mathbf{p} = \mathcal{E}\{\mathbf{h}_r \mathbf{h}_r^T\}. \tag{30.10}$$

The mean $\hat{\mathbf{h}}$ then represents a systematic (i.e., deterministic) component of the uncertain echo path $\mathbf{h}$, whereas the residual $\mathbf{h}_r$ is its truly unpredictable (i.e., zero-mean) component. The signal model as presented here clearly fits the practical applications of acoustic echo control in which the echo path is usually unknown, but the echo path input is in fact known to the system. The implications of the echo path uncertainty model will developed in the following sections in conjunction with the derivation of optimum filters.

Author's personal copy

4.30.2.2 Generalized Wiener filter architecture

We initiate the actual derivation of the optimal AEC by a formal definition of the desired quality of near-end speech reconstruction in the form of an objective function. On the one hand, full-duplex operation of the hands-free system in Figure 30.1 ideally requires strong attenuation of the acoustic echo signal $d(n)$ by the acoustic echo controller. On the other hand, the echo attenuation is subject to the undistorted reproduction of the desired signal $s(n)$ at the system output. Mathematically, this conflict can be expressed as a statistical optimization problem which aims, for example, at the minimum mean-square error (MMSE) between $s(n)$ and $\hat{s}(n)$:

$$\epsilon^2 = \mathcal{E}\left\{\left(s(n) - \hat{s}(n)\right)^2\right\} \rightarrow \min. \tag{30.11}$$

In order to facilitate the computation of the system output $\hat{s}(n)$, we formulate the echo control problem as a general unconstrained linear filtering problem, where the output signal $\hat{s}(n)$ is obtained as a linear combination of the available input signals $x(n)$ and $y(n)$:

$$\hat{s}(n) = w'_{2,n} * y(n) + w'_{1,n} * x(n) \tag{30.12}$$

$$= w_{2,n} * \left[y(n) - w_{1,n} * x(n)\right]. \tag{30.13}$$

Mathematically, the filter structures in (30.12) and (30.13) are equivalent, since they can be uniquely transformed into each other. In principle, we can either optimize the linear filters $w'_{1,n}$ and $w'_{2,n}$, or alternatively $w_{1,n}$ and $w_{2,n}$. It turns out, however, that the solution for the filter structure in (30.13) is somewhat simpler and more intuitive. To be in line with the common literature on acoustic echo control, we will then refer to $w_{1,n}$ as the echo canceler and to $w_{2,n}$ as the postfilter for residual echo suppression. The formulation as an unconstrained, possibly IIR filtering problem emphasizes that the echo control problem is not undermodeled by a strict limitation of the adaptive filter length here. That implies that the echo canceler $w_{1,n}$ can entirely cover the span of the linear echo path h_n. Since $w_{1,n} = h_n$ would clearly eliminate the echo, it will be interesting to clarify the role of the postfilter $w_{2,n}$ in this seemingly simple configuration.

The filter structure in (30.13) requires the implementation of two consecutive convolutions, i.e., the convolution of $x(n)$ with $w_{1,n}$, and the subsequent convolution of the difference $e(n) = y(n) - w_{1,n} * x(n)$ with the filter $w_{2,n}$. In vector notation, that means that we have to provide a vector

$$\mathbf{e}(n) = \left[e(n), e(n-1), \ldots, e(n-N_2)\right]^T \tag{30.14}$$

of input samples for the second convolution and thus a vector of output samples from the first convolution. That in turn requires the definition of an input signal matrix

$$\mathbf{X}(n) = \left[\mathbf{x}(n), \mathbf{x}(n-1), \ldots, \mathbf{x}(n-N_2)\right]^T, \tag{30.15}$$

which then allows us to write the AEC output according to (30.13) as

$$\hat{s}(n) = \mathbf{w}_2^T \mathbf{e}(n)$$

$$= \mathbf{w}_2^T \left[\mathbf{y}(n) - \mathbf{X}(n)\mathbf{w}_1\right] \tag{30.16}$$

$$= \left[\mathbf{y}^T(n) - \mathbf{w}_1^T \mathbf{X}^T(n)\right]\mathbf{w}_2, \tag{30.17}$$

Author's personal copy

where

$$\mathbf{y}(n) = \left[y(n), y(n-1), \ldots, y(n-N_2)\right]^T \tag{30.18}$$

is a vector of most recent microphone samples and

$$\mathbf{w}_1 = \left[w_{1,0}, w_{1,1}, \ldots, w_{1,N_1}\right]^T \tag{30.19}$$

$$\mathbf{w}_2 = \left[w_{2,0}, w_{2,1}, \ldots, w_{2,N_2}\right]^T \tag{30.20}$$

are the coefficient vectors of length N_1 and length N_2 of echo canceler and postfilter, respectively. The two forms in (30.16) and (30.17) are equivalent.

We further exploit the available vector-matrix notation to write the microphone signal vector $\mathbf{y}(n)$ as

$$\mathbf{y}(n) = \mathbf{s}(n) + \mathbf{X}(n)\mathbf{h} \tag{30.21}$$

in accordance with the original convolutive signal model in (30.4).

We then proceed by substituting the two-stage linear filter structure (30.13) into (30.11) and by computing the partial derivatives of the mean-square error ϵ^2 with respect to the coefficients $\mathbf{w}_1$ and $\mathbf{w}_2$. The following expressions for the derivatives are obtained using not more than the previously made assumptions of a deterministic input signal $\mathbf{X}(n)$ and independent near-end speech $s(n)$ with zero mean:

$$\begin{aligned}
\frac{\partial \epsilon^2}{\partial \mathbf{w}_1} &= -2\mathcal{E}\left\{\left(s(n) - \hat{s}(n)\right)\frac{\partial \hat{s}(n)}{\partial \mathbf{w}_1}\right\} \\
&= 2\mathcal{E}\left\{\hat{s}(n)\right\}\mathbf{X}^T(n)\mathbf{w}_2 \\
&= 2\left[\mathcal{E}\left\{\mathbf{y}^T(n)\right\}\mathbf{w}_2 - \mathbf{w}_1^T\mathbf{X}^T(n)\mathbf{w}_2\right]\mathbf{X}^T(n)\mathbf{w}_2,
\end{aligned} \tag{30.22}$$

$$\begin{aligned}
\frac{\partial \epsilon^2}{\partial \mathbf{w}_2} &= -2\mathcal{E}\left\{\left(s(n) - \hat{s}(n)\right)\frac{\partial \hat{s}(n)}{\partial \mathbf{w}_2}\right\} \\
&= -2\mathcal{E}\left\{\left(s(n) - \hat{s}(n)\right)\left(\mathbf{y}(n) - \mathbf{X}(n)\mathbf{w}_1\right)\right\} \\
&= -2\mathcal{E}\left\{\hat{s}(n)\right\}\mathbf{X}(n)\mathbf{w}_1 - 2\mathcal{E}\left\{\mathbf{y}(n)\left(s(n) - \hat{s}(n)\right)\right\}.
\end{aligned} \tag{30.23}$$

Here, the reason for the remaining time index n after the evaluation of the statistical expectation lies in the deterministic nature of $\mathbf{X}(n)$.

In the next step of our derivation, we make use of the linear signal model in (30.21) and find that $\mathcal{E}\left\{\mathbf{y}^T(n)\right\} = \mathcal{E}\left\{\mathbf{h}^T\right\}\mathbf{X}^T(n)$. Now it can be easily seen that $\partial\epsilon^2/\partial\mathbf{w}_1 = \mathbf{0}$ by choosing the optimum echo canceler coefficients as

$$\mathbf{w}_1 = \mathcal{E}\{\mathbf{h}\} = \hat{\mathbf{h}}. \tag{30.24}$$

In order to find the postfilter $\mathbf{w}_2$ that satisfies $\partial\epsilon^2/\partial\mathbf{w}_2 = \mathbf{0}$, we consider the last line of Eq. (30.23) and initially note that $\mathcal{E}\left\{\hat{s}(n)\right\}$ again vanishes due to the choice $\mathbf{w}_1 = \hat{\mathbf{h}}$ for the echo canceler, as before in (30.22). The second part of (30.23) can be expanded by inserting the system output $\hat{s}(n)$ according to (30.17) and the signal model for $\mathbf{y}(n)$ as shown in (30.21):

$$\begin{aligned}
\mathcal{E}\left\{\mathbf{y}(n)\left(s(n) - \hat{s}(n)\right)\right\} &= \mathcal{E}\left\{\mathbf{y}(n)\left[s(n) - \left(\mathbf{y}^T(n) - \mathbf{w}_1^T\mathbf{X}^T(n)\right)\mathbf{w}_2\right]\right\} \\
&= \mathcal{E}\left\{\left[\mathbf{s}(n) + \mathbf{X}(n)\mathbf{h}\right]\left[s(n) - \left(\mathbf{s}^T(n) + \mathbf{h}^T\mathbf{X}^T(n) - \mathbf{w}_1^T\mathbf{X}^T(n)\right)\mathbf{w}_2\right]\right\}.
\end{aligned}$$

Author's personal copy

By then utilizing the uncertainty model $\mathbf{h} = \hat{\mathbf{h}} + \mathbf{h}_r$ of the echo path as shown in (30.9), and invoking the optimum solution $\mathbf{w}_1 = \hat{\mathbf{h}}$ in (30.24), and finally the independence of $\mathbf{h}_r$ and $s(n)$, we arrive at the following equation (consisting of deterministic and statistical terms) for the optimum filter $\mathbf{w}_2$:

$$\begin{aligned}\mathcal{E}\left\{\mathbf{y}(n)\left(s(n)-\hat{s}(n)\right)\right\} &= \mathcal{E}\left\{\left[\mathbf{s}(n)+\mathbf{X}(n)(\hat{\mathbf{h}}+\mathbf{h}_r)\right]\left[s(n)-\left(\mathbf{s}^T(n)+\mathbf{h}_r^T\mathbf{X}^T(n)\right)\mathbf{w}_2\right]\right\}\\ &= \mathcal{E}\left\{\mathbf{s}(n)s(n)\right\}-\mathcal{E}\left\{\mathbf{s}(n)\mathbf{s}^T(n)\mathbf{w}_2\right\}-\mathcal{E}\left\{\mathbf{X}(n)\mathbf{h}_r\mathbf{h}_r^T\mathbf{X}^T(n)\mathbf{w}_2\right\}\\ &= \mathbf{r}_s-\mathbf{R}_s\mathbf{w}_2-\mathbf{X}(n)\mathbf{p}\mathbf{X}^T(n)\mathbf{w}_2=\mathbf{0},\end{aligned} \tag{30.25}$$

where $\mathbf{r}_s = \mathcal{E}\{\mathbf{s}(n)s(n)\}$is the autocorrelation vector of the near-end speech being equal to the first column of $\mathbf{R}_s$. From here, we can easily solve for the optimum postfilter $\mathbf{w}_2$ by rearranging the vector-matrix equation to

$$\mathbf{w}_2 = \left(\mathbf{R}_s+\mathbf{X}(n)\mathbf{p}\mathbf{X}^T(n)\right)^{-1}\mathbf{r}_s. \tag{30.26}$$

We note that the result for $\mathbf{w}_2$ has the structure of a Wiener filter to perform noise reduction on the signal

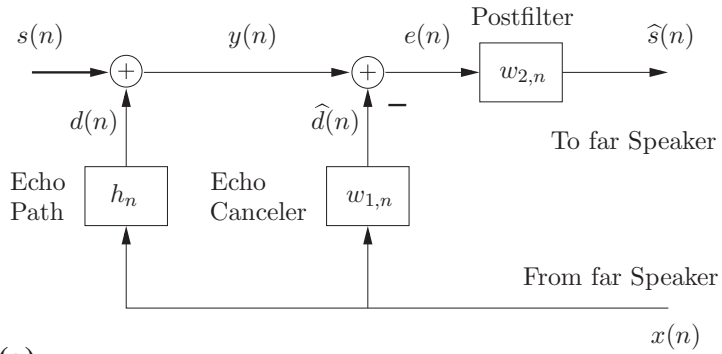
$$e(n) = y(n)-\hat{\mathbf{h}}^T\mathbf{x}(n) = s(n)+\mathbf{h}_r^T\mathbf{x}(n), \tag{30.27}$$

with $\mathbf{h}_r^T\mathbf{x}(n)$ being the effective noise (here: The residual echo) and $s(n)$ the desired signal. In this interpretation, the compound quantity $\mathbf{X}(n)\mathbf{p}\mathbf{X}^T(n)$ in (30.26) can be considered as the noise autocorrelation matrix. It is in fact obtained as a weighted short-term autocorrelation $\mathbf{X}(n)\mathbf{X}^T(n)$ of the input signal $x(n)$, where the weighting matrix $\mathbf{p}$ serves as a statistical descriptor of the uncertain residual echo transmission system $\mathbf{h}_r = \mathbf{h}-\hat{\mathbf{h}}$. Commutativity of signals and systems further allows for the equivalent and yet more intuitive understanding of the random quantity $\mathbf{h}_r$ as the input of a deterministic transmission system described by the known vector $\mathbf{x}(n)$. The latter interpretation right away explains the covariance $\mathbf{X}(n)\mathbf{p}\mathbf{X}^T(n)$ of the output of the convolution $\mathbf{h}_r^T\mathbf{x}(n) = \mathbf{x}^T(n)\mathbf{h}_r$.

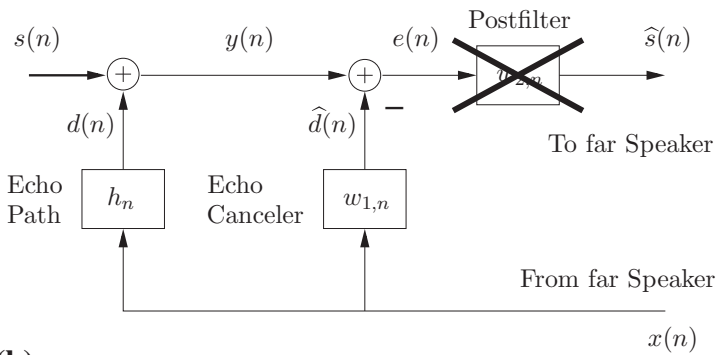
In our derivation, echo canceler and postfilter have been deduced jointly from the MMSE criterion and, therefore, we will refer to the combination of (30.24) and (30.26) as the *generalized Wiener solution* for acoustic echo control. The optimization was based on a signal model which consists of an uncertain linear echo path with a deterministic input signal. We assume that this signal model greatly fits the practical applications of acoustic echo control in which the echo path is usually unknown and the echo path input is in fact measurable. A block diagram of the resulting optimal filter structure immediately follows from Eq. (30.13). It is depicted in Figure 30.3a.

Regarding the practical implementation of the algorithms, however, we have to mention that the assumption of stationarity of the near-end speech $s(n)$ is of course not realistic. As usual in speech and audio processing, the optimum filters thus have to be updated at a time-constant of 10–30 ms on the basis of short-term stationary signal frames. Moreover, the calculation of the postfilter in (30.26) requires efficient algorithms for matrix inversion, such as the generalized Levinson algorithm [89], or we can alternatively approach efficient solutions in the frequency-domain [27]. The digital filtering as such, according to (30.16), can be realized by fast convolution in the DFT domain or by direct convolution in the time-domain [90].

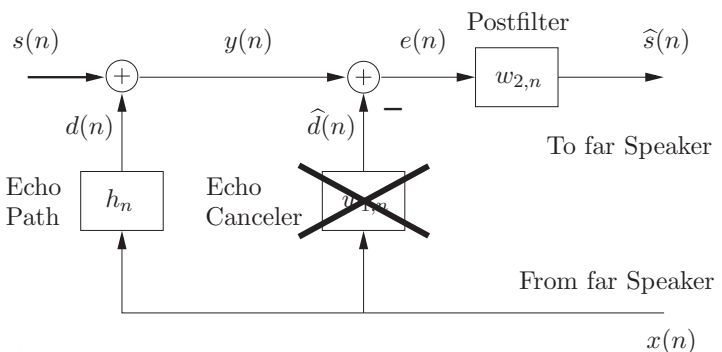
Author's personal copy



(a) Generalized Wiener solution for the uncertain linear echo path model with non-zero mean $\mathcal{E}\{\mathbf{h}\}$ and non-zero covariance $\mathbf{p}$.



(b) Classical Wiener solution for the deterministic echo path model without uncertainty, i.e., $\mathcal{E}\{\mathbf{h}\} = \mathbf{h}$ and $\mathbf{p} = 0$.



(c) Wiener solution for an unpredictable echo path with zero-mean, i.e., $\mathcal{E}\{\mathbf{h}\} = \mathbf{0}$ and $\mathbf{p} = \mathbf{R}_h = \mathcal{E}\{\mathbf{h}\mathbf{h}^T\}$.

FIGURE 30.3

The generalized Wiener solution and important special cases.

Author's personal copy

4.30.2.3 General and special cases

Before moving on to adaptive algorithms dedicated for the generalized Wiener filter, we shall clarify its significance in comparison with two special cases of it. We will demonstrate that the general solution and its better known special cases correspond to different filter structures, mainly differing in the way they utilize *a priori* information in form of the mean and the covariance of the acoustic echo path, depending on the application at hand.

4.30.2.3.1 General statistical case

For the convenience of the reader, we first repeat the two optimum filters:

$$\mathbf{w}_1 = \mathcal{E}\{\mathbf{h}\} \tag{30.28}$$

$$\mathbf{w}_2 = \left(\mathbf{R}_s + \mathbf{X}(n)\mathbf{p}\mathbf{X}^T(n)\right)^{-1}\mathbf{r}_s. \tag{30.29}$$

It can be seen that the generalized Wiener filter fully takes the statistical properties of the echo path $\mathbf{h}$ into account. More specifically, the mean and covariance of the echo path are utilized separately to determine the optimum echo canceler and postfilter coefficients, respectively. The optimum echo canceler $\mathbf{w}_1 = \mathcal{E}\{\mathbf{h}\} = \hat{\mathbf{h}}$ creates a replica of the echo components which are due to the systematic component of the echo path, i.e., $\hat{d}(n) = \hat{\mathbf{h}}^T\mathbf{x}(n)$. The echo subtraction according to Figure 30.3a therefore results in the error signal $e(n) = y(n) - \hat{d}(n) = s(n) + \mathbf{h}_r^T\mathbf{x}(n)$, comprising the desired signal $s(n)$ plus undesired residual echo $\mathbf{h}_r^T\mathbf{x}(n)$. The signal $e(n)$ is then postfiltered with coefficients according to (30.29), thereby taking the echo path covariance $\mathbf{p}$ in conjunction with the input signal $\mathbf{X}(n)$ into account.

It turns out in practice that this general two-filter solution is very much suitable to achieve acoustic echo control in hands-free communication systems, such as in car hands-free telephones to be described in Section 4.30.6.1. In many other applications, too, it is indeed feasible to determine a (possibly time-varying) systematic component $\hat{\mathbf{h}}$ of the echo path $\mathbf{h}$ using the acoustic system identification approach, e.g., [2,35]. Nevertheless, some degree of uncertainty $\mathbf{p}$ about the echo path always remains and, therefore, the postfilter is an indispensable component of advanced hands-free telephones.

4.30.2.3.2 Deterministic case

In special cases without uncertainty about the echo path at all, i.e., if $\mathcal{E}\{\mathbf{h}\} = \mathbf{h}$ and thus $\mathbf{p} = 0$, our general solution degenerates to

$$\mathbf{w}_1 = \mathbf{h} \tag{30.30}$$

$$\mathbf{w}_2 = \mathbf{u}, \tag{30.31}$$

where $\mathbf{u} = [1, 0, 0, \ldots, 0]^T$ denotes the unit vector of length N_2. Here, the optimum echo canceler $\mathbf{w}_1$ is an ideal copy of the true echo path $\mathbf{h}$ and the perfectly echo-canceled error signal $e(n) = y(n) - \mathbf{h}^T\mathbf{x}(n) = s(n)$ will naturally pass the postfilter $\mathbf{w}_2$ unprocessed. This procedure is commonly understood as the Wiener solution for acoustic echo cancellation [2,10] and it is indeed optimal for a deterministic echo path $\mathbf{h}$. The filter structure corresponding to this special case is illustrated in Figure 30.3b.

Author's personal copy

The purely deterministic echo path model assumed here is often not the best choice in real systems. It may require extremely sophisticated adaptive filters and control mechanisms to let the echo canceler coefficients follow a true time-varying echo path with sufficient accuracy. It is obvious that especially in noisy and time-varying acoustic environments, as for example in vehicles, the true echo path cannot be determined exactly at all times. As a consequence, the deterministic strategy may not deliver sufficient echo attenuation, i.e., residual echo can appear at the system output. The situation can be different in applications of network or line echo cancellation. Here, the echo path coefficients can often be measured with sufficient accuracy during call setup and they are not expected to change significantly during conversation. A postfilter for residual echo suppression is then not needed.

4.30.2.3.3 *Zero-mean case*

When no systematic information is available about the echo path at all, i.e., if $\mathcal{E}\{\mathbf{h}\} = \mathbf{0}$ and thus $\mathbf{p} = \mathbf{R}_h = \mathcal{E}\{\mathbf{h}\mathbf{h}^T\}$, we obtain another special case of the generalized Wiener solution:

$$\mathbf{w}_1 = 0, \tag{30.32}$$

$$\mathbf{w}_2 = \left(\mathbf{R}_s + \mathbf{X}(n)\mathbf{R}_h\mathbf{X}^T(n)\right)^{-1}\mathbf{r}_s. \tag{30.33}$$

Obviously, the generalized Wiener filter degenerates to an MMSE equalizer in sending direction of the communication system, i.e., the responsibility for acoustic echo suppression is entirely with the statistical postfilter. The corresponding block diagram in Figure 30.3c clearly reminds ourselves to the methodology of background noise suppression. The difference, however, is that the known far-end speech $\mathbf{X}(n)$ is taken into account to calculate the noise PSD in conjunction with the echo path covariance $\mathbf{R}_h$.

The practical relevance of the zero-mean case is given by applications in which it turns out difficult to determine a systematic component of the echo path at all. This situation can appear in extremely time-varying and noisy systems, e.g., teleconferencing equipment in reverberant environments or hands-free accessories with unstable microphone and loudspeaker placement. Mobile devices may simply not provide the computational resources for a sophisticated echo path estimator. At the same time, it should be noted that the desired speech at the system output can be distorted and the background noise of the near-end can be modulated by the presence of the far-end speech signal. Section 4.30.3 is dedicated to the zero-mean case and its implications.

4.30.2.4 Dynamical echo path modeling

In almost all practical applications of acoustic echo control, such as car hands-free systems, teleconferencing equipment, mobile phones, and speech dialog systems, we face a time-varying acoustic echo path $\mathbf{h}(n)$. Depending on the application, the degree of change can be more or less pronounced. In reality, we further have to expect sometimes quite abrupt changes and sometimes almost static behavior of the echo path. On average, the acoustic echo path is certainly not standing still, but also not changing arbitrarily fast, i.e., the dynamical process is governed by finite bandwidth.

In order to exploit the average smooth dynamical nature of the echo path in adaptive algorithm development, we shall now refine our previous uncertainty model of the echo path in (30.8)–(30.10) .

Author's personal copy

In this respect, a particularly convenient stochastic model for time-varying systems $\mathbf{h}(n)$ is the first-order recursive Markov chain [10], i.e.,

$$\mathbf{h}(n+1) = a \cdot \mathbf{h}(n) + \Delta\mathbf{h}(n), \tag{30.34}$$

where two consecutive realizations at times n and $n+1$ are related to each other by the transition coefficient $0 \leq a \leq 1$ and the independent process noise quantity $\Delta\mathbf{h}(n)$ with zero mean and covariance matrix $\mathbf{R}_\Delta = \mathrm{E}\{\Delta\mathbf{h}(n)\Delta\mathbf{h}^T(n)\}$. The Markov model therefore represents dynamic behavior in which the state $\mathbf{h}(n)$ gradually changes into an unpredictable direction—very much in agreement with the nature of time-varying impulse responses in realistic acoustic environments.

Clearly, the Markov model will serve only as a simplified model of the real world situation. However, it brings along two major relationships with real time-varying systems by, firstly, restricting the bandwidth of change according to the transition factor "a" and, secondly, providing an element of uncertainty through the process noise $\Delta\mathbf{h}(n)$. In order to describe different degrees of variability, we might intuitively adjust the transition factor or the process noise covariance of the model. However, to be sure about the consequences, we shall more formally consider some properties of the Markov model in conjunction with the acoustic echo control purpose:

- With the process noise $\Delta\mathbf{h}(n)$ assumed to be zero-mean and stationary, where the latter is reflected explicitly by the time-invariant process noise covariance $\mathbf{R}_\Delta$, and by considering the linear time-invariant system in (30.34), the echo path $\mathbf{h}(n)$ is immediately recognized as a zero-mean and stationary random process, too, and described by the time-invariant covariance $\mathbf{R}_h = \mathrm{E}\{\mathbf{h}(n)\mathbf{h}^T(n)\}$. Such properties of $\mathbf{h}(n)$ are well in agreement with the acoustic echo control application. Here, the zero mean, $\mathrm{E}\{\mathbf{h}(n)\} = 0$, in fact represents the average over all possible echo paths when no suitable *a priori* information is available about the electroacoustic environment—including gain and phase of loudspeaker and microphone amplifiers, the exact physical distance between loudspeaker and microphone, and the room characteristics. The time-invariance of the echo path covariance $\mathbf{R}_h$ further expresses the persistence of the acoustic echo path impulse response $\mathbf{h}(n)$, independent of the highly nonstationary echo path input $\mathbf{x}(n)$.
- By applying square expectation on both sides of (30.34), and by exploiting stationarity of $\mathbf{h}(n)$, i.e., $\mathbf{R}_h = \mathrm{E}\{\mathbf{h}(n+1)\mathbf{h}^T(n+1)\} = \mathrm{E}\{\mathbf{h}(n)\mathbf{h}^T(n)\}$, and utilizing the independence of $\Delta\mathbf{h}(n)$, we can evaluate the echo path covariance as follows:

$$\mathbf{R}_h = a^2\mathbf{R}_h + \mathbf{R}_\Delta. \tag{30.35}$$

  This result can be rearranged to obtain an interesting proportionality between the covariances of echo path changes and echo path:

$$\mathbf{R}_\Delta = (1 - a^2)\mathbf{R}_h. \tag{30.36}$$

  From this relationship, we learn that we cannot choose the transition factor "a" and the process noise covariance $\mathbf{R}_\Delta$ of the Markov model in (30.34) independently, since the echo path covariance $\mathbf{R}_h$ is typically given as a somewhat fixed and persistent quantity, despite the possibly changing acoustic impulse response $\mathbf{h}(n)$. Moreover, the relation in (30.36) can even be very useful to determine an unknown covariance of the echo path changes, $\mathbf{R}_\Delta$, from an estimated or *a priori* known echo path covariance $\mathbf{R}_h$.

Author's personal copy

Finally, the state Eq. (30.34) and the linear observation model (30.4) can be formally combined into a general stochastic state-space model of the unknown echo path state $\mathbf{h}(n)$. For convenience and to include the time-varying echo path into the observation equation, both models are repeated here together:

$$\mathbf{h}(n+1) = a \cdot \mathbf{h}(n) + \Delta\mathbf{h}(n), \tag{30.37}$$

$$y(n) = s(n) + \mathbf{x}^T(n)\mathbf{h}(n). \tag{30.38}$$

In summary, the echo path state equation is governed by the independent process noise $\Delta\mathbf{h}(n)$ with covariance $\mathbf{R}_\Delta$. The resulting state $\mathbf{h}(n)$ is then observed through the microphone signal $y(n)$ in the presence of near-end speech $s(n)$ which acts as independent observation noise with covariance σ_s^2. This statement of the AEC problem will now lead us to the utilization of powerful state estimators from control theory to deduce contained and efficient adaptive algorithms for acoustic echo control.

4.30.2.5 Adaptive algorithms

The previous section has argued for modeling the *a priori* echo path mean $\mathcal{E}\{\mathbf{h}(n)\} = \hat{\mathbf{h}}$ as zero. The optimum echo canceler in (30.28) would thus degenerate to $\mathbf{w}_1(n) = \hat{\mathbf{h}} = 0$, practically meaning that the generalized Wiener solution turns into the MMSE equalizer in (30.33). In order to exploit the full-featured Wiener solution, we have to resolve at least partly the uncertainty about the echo path $\mathbf{h}(n)$. This can be done by recasting the uncertainty framework in a way that incorporates the observations $y(n)$ into the stochastic echo path model. Mathematically, this can be accomplished by defining the *conditional* mean and covariance of the echo path at time n, given the observations $y(n)$ up to and including time $n-1$:[2]

$$\hat{\mathbf{h}}(n) = \mathcal{E}\{\mathbf{h}(n)|y(n-1), y(n-2), \ldots, y(0)\}, \tag{30.39}$$

$$\mathbf{h}_r(n) = \mathbf{h}(n) - \hat{\mathbf{h}}(n), \tag{30.40}$$

$$\mathbf{p}(n) = \mathcal{E}\{\mathbf{h}_r(n)\mathbf{h}_r^T(n)\}. \tag{30.41}$$

This data driven uncertainty model of the echo path basically replaces the *a priori* echo path model in (30.8)–(30.10). In place of the *a priori* mean $\hat{\mathbf{h}}$, we now have the time-varying conditional mean $\hat{\mathbf{h}}(n)$. The former residual $\mathbf{h}_r$ in (30.9) accordingly has been replaced by its time-varying counterpart $\mathbf{h}_r(n)$, representing the misalignment between the now conditional echo path mean and the true echo path $\mathbf{h}(n)$. Moreover, the former *a priori* echo path covariance $\mathbf{p}$ has turned into the now time-varying echo path covariance $\mathbf{p}(n)$ based on the conditional mean. Structurally, the definitions (30.39)–(30.41) are fully consistent with the previous ones in (30.8)–(30.10). Therefore, the conditional mean $\hat{\mathbf{h}}(n)$ can serve as an optimum filter for subtractive echo cancellation in (30.28) along with the conditional covariance $\mathbf{p}(n)$ in place of $\mathbf{p}$ in the postfilter Eq. (30.29).

[2] In the application of acoustic echo control, the most recent data $y(n)$ at time n is usually not included into the estimation of the acoustic echo path $\mathbf{h}(n)$ at time n, e.g., consider the LMS, APA, and RLS family of adaptive algorithms [2]. This has the practical advantage that the estimated echo path is completely determined already at time $n-1$ and can be employed for echo cancellation immediately when the input data $y(n)$ is available. In the language of state-space modeling and estimation, the conditional mean $\hat{\mathbf{h}}(n)$ in (30.40) is called the *a priori* estimate of the state $\mathbf{h}(n)$. An *a posteriori* estimate $\hat{\mathbf{h}}^+(n)$ which is a refinement of $\hat{\mathbf{h}}(n)$ based on the current data $y(n)$ could be defined as well.

Author's personal copy

The computation of the conditional expectation in (30.39) and the corresponding covariance in (30.41), subject to the state-space model in (30.37) and (30.38), is a well understood problem. The adequate mathematical instrument for the solution is the statistical Kalman filter. In literature, we find several principal interpretations of it. In [10], the Kalman filter is derived as the *linear* MMSE estimator of the state of a linear dynamical system. In [91], the Kalman filter is developed as the MMSE state estimator under the assumption of *Gaussianity* of process noise and observation noise. A very intuitive presentation of the Kalman filter equations can be found for example in [92]. The original work of Kalman is documented in [93]. Independent of the particular interpretation, the Kalman filter delivers at least a good approximation of the conditional mean $\hat{\mathbf{h}}(n)$ and the corresponding estimation error covariance $\mathbf{p}(n)$ with respect to the unknown parameter vector $\mathbf{h}(n)$. The algorithm consists of the following set of recursive and iteratively coupled matrix equations:

$$\hat{\mathbf{h}}(n+1) = a \cdot \hat{\mathbf{h}}^{+}(n), \tag{30.42}$$

$$\mathbf{p}(n+1) = a^2 \cdot \mathbf{p}^{+}(n) + \mathbf{R}_{\Delta}, \tag{30.43}$$

$$\hat{\mathbf{h}}^{+}(n) = \hat{\mathbf{h}}(n) + \mathbf{k}(n)(y(n) - \mathbf{x}^T(n)\hat{\mathbf{h}}(n)), \tag{30.44}$$

$$\mathbf{p}^{+}(n) = (\mathbf{I} - \mathbf{k}(n)\mathbf{x}^T(n))\mathbf{p}(n), \tag{30.45}$$

$$\mathbf{k}(n) = \mathbf{p}(n)\mathbf{x}(n)\left(\mathbf{x}^T(n)\mathbf{p}(n)\mathbf{x}(n) + \sigma_s^2(n)\right)^{-1}. \tag{30.46}$$

Equations (30.42) and (30.44) recursively determine the conditional mean $\hat{\mathbf{h}}(n)$ in a prediction-correction style. In doing so, the formulas utilize the Kalman gain $\mathbf{k}(n)$ from (30.46) as a weight vector which essentially depends on the state error covariance $\mathbf{p}(n)$. The latter is again determined recursively through Eqs. (30.43) and (30.45) of the Kalman filter.

The Kalman gain $\mathbf{k}(n)$ can be considered as an intelligent adaptive stepsize parameter in the recursive learning procedure for the conditional echo path mean $\hat{\mathbf{h}}(n)$, i.e., $\mathbf{k}(n)$ basically upgrades the role of the fixed stepsize μ in the LMS algorithm [2,10]. Through the Kalman gain, the model-based "system distance" $\mathbf{p}(n)$ between the true and the estimated acoustic system interacts with the prediction-correction procedure for the estimation of $\hat{\mathbf{h}}(n)$. In this way, Kalman filtering can be understood as the ever sought unification of linear adaptive filtering and adaptation control. After all, the Kalman filter differs from LMS and RLS by its inherent stability [10], i.e., it does not require additional control mechanisms (e.g., the double-talk detection) in order to achieve fast and yet robust adaptation in time-varying and noisy acoustic environments.

So far, the Kalman filter has been employed for acoustic system identification hardly ever. This can be attributed to its high computational load and to the risk for numerical instability in the case of higher-order adaptive filters [10]. Furthermore, a comprehensive signal model for the Kalman filter, particularly the availability of observation and process noise covariances for the acoustic state-space model in (30.38) and (30.37), seemed to be out of sight [2].

In order to tame the exact Kalman filter, we briefly outline the procedure as described in [94]. At first, we replace the matrix quantity $\mathbf{k}(n)\mathbf{x}^T(n)$ in (30.45) with the inner vector product $\mathbf{x}^T(n)\mathbf{k}(n)/N_1$. This seemingly brutal simplification can be well justified in the case of broadband input $x(n)$, since the recursively smoothed matrix quantity $\mathbf{k}(n)\mathbf{x}^T(n)$ resembles the near-diagonal correlation matrix of $x(n)$. Provided that we specify a diagonal process noise covariance $\mathbf{R}_{\Delta} = \sigma_{\Delta}^2\mathbf{I}$, the state error covariance matrix $\mathbf{p}(n)$ then can be treated as a scalar $p(n)$ without further assumption or approximation, as seen

Author's personal copy

from (30.43) and (30.45). The normalization by factor N_1 in the former replacement $\mathbf{x}^T(n)\mathbf{k}(n)/N_1$ achieves appropriate scaling after the substitution. Due to the broadband rationale behind these rearrangements, the resulting algorithm is termed *broadband Kalman filter*:

$$\hat{\mathbf{h}}(n+1) = a \cdot \hat{\mathbf{h}}^+(n), \tag{30.47}$$

$$p(n+1) = a^2 \cdot p^+(n) + \sigma_\Delta^2, \tag{30.48}$$

$$e(n) = y(n) - \mathbf{x}^T(n)\hat{\mathbf{h}}(n), \tag{30.49}$$

$$\hat{\mathbf{h}}^+(n) = \hat{\mathbf{h}}(n) + \mathbf{k}(n)e(n), \tag{30.50}$$

$$p^+(n) = \left(1 - \mathbf{x}^T(n)\mathbf{k}(n)/N_1\right) p(n), \tag{30.51}$$

$$\mathbf{k}(n) = p(n)\mathbf{x}(n)\left(p(n)\mathbf{x}^T(n)\mathbf{x}(n) + \sigma_s^2(n)\right)^{-1}. \tag{30.52}$$

By the simplifications introduced here, naturally, the presented algorithm loses its decorrelation ability regarding the input signal $x(n)$ if non-white input is processed. However, all the structural support to handle the estimation of time-varying unknown systems $\mathbf{h}(n)$ in the continuous presence of observation noise $s(n)$, with possibly time-varying level $\sigma_s^2(n)$, is fully preserved in the *broadband Kalman filter*. Moreover, we have at the same time gained considerable numerical efficiency by reducing the dimension of the original state error covariance $\mathbf{p}(n)$ from matrix to scalar.

Next, we mention an opportunity to resolve a possible uncertainty regarding the time-varying observation noise power $\sigma_s^2(n)$ in the Kalman gain (30.52), because this quantity is indispensable for the operation of the Kalman filter. Unfortunately, the corresponding signal $s(n)$ is not available explicitly for the calculation of sample covariances, but the error signal $e(n)$ in (30.49) represents at least a good estimate of the observation noise signal $s(n)$ in case of successful state estimation. Thus, we can approximate $\sigma_s^2(n) \approx \sigma_e^2(n)$ and then obtain the error signal power $\sigma_e^2(n)$, e.g., by recursive averaging of the explicitly available square error $e^2(n)$.

The scalar process noise covariance parameter σ_Δ^2 required in (30.48) can be specified as $\sigma_\Delta^2 = (1-a^2)\mathcal{E}\{\mathbf{h}^T(n)\mathbf{h}(n)\}/N_1$, where $\mathcal{E}\{\mathbf{h}^T(n)\mathbf{h}(n)\}$ denotes an expectation of the echo path norm. This formula is in line with (30.36) by again reducing the process noise covariance from matrix to scalar dimension as done already in the derivation of the broadband Kalman filter, $\mathbf{R}_\Delta = \mathbf{I}\sigma_\Delta^2$, and similarly $\mathbf{R}_h = \mathbf{I}\sigma_h^2 = \mathbf{I}\mathcal{E}\{\mathbf{h}^T(n)\mathbf{h}(n)\}/N_1$. The transition parameter "a" has to be determined appropriately for the application at hand.

Substituting (30.50) and (30.52) into (30.47), while assuming low near-end speech and near-end noise, i.e., $\sigma_s^2(n) \to 0$, the broadband Kalman filter reveals structural equivalence with the NLMS algorithm [10], except for the leaky factor a which appears in the update equation (instead of a stepsize factor as for the pure NLMS algorithm):

$$\hat{\mathbf{h}}(n+1) = a \cdot \left(\hat{\mathbf{h}}(n) + \mathbf{k}(n)e(n)\right), \tag{30.53}$$

$$\mathbf{k}(n) = \left(\mathbf{x}^T(n)\mathbf{x}(n)\right)^{-1}\mathbf{x}(n), \tag{30.54}$$

$$e(n) = y(n) - \mathbf{x}^T(n)\hat{\mathbf{h}}(n). \tag{30.55}$$

Author's personal copy

The resulting NLMS algorithm proves the numerical efficiency and robustness obtained through the simplification of the exact Kalman filter. The *broadband Kalman filter*, still with $\sigma_s^2(n) \neq 0$, then in fact represents an excellent compromise in terms of adaptive performance and structural simplicity between exact Kalman filters and very popular LMS-type adaptive algorithms for acoustic system identification. In contrast to our top-down justification of the broadband Kalman filter, a model-based bottom-up generalization of the NLMS algorithm leads to a similar algorithm [95,96].

Resuming to more general considerations, we finally want to mention that and alternative and preferred realization of adaptive algorithms for joint conditional mean and covariance estimation has been presented in the literature. In [27], a *block frequency-domain adaptive Kalman filter* and the underlying state-space model in the DFT domain have been suggested. It was demonstrated that the aforementioned drawbacks of the exact Kalman filter in time-domain can be circumvented very efficiently by diagonalization—through minor approximation—of the matrix algebra in the DFT domain. Intuitive approaches for process and observation noise covariance learning were suggested, similar to the ones presented right above, and the impact of model mismatch between the underlying states-space model of the Kalman filter and real world dynamics has been reported to be "graceful." In [97], the DFT-domain adaptive algorithm was then reformulated as a *state-space frequency-domain adaptive filter* (SSFDAF) in the context of acoustic echo control. In [98], the intuitive way of covariance learning was confirmed in the maximum-likelihood sense and the superiority of state-space frequency-domain adaptive filtering in comparison to traditional frequency-domain adaptive filtering was demonstrated.

The block frequency-domain adaptive Kalman filter is not limited to, but especially suits the adaptation of a frequency-domain representation of the generalized Wiener filter as shown in [27]. This concept and implementation will be used as the basis for the presentation of numerical results in the application of car hands-free systems in Section 4.30.6.1.

4.30.3 Echo suppression

An alternative way to prevent acoustic echoes is the use of an acoustic echo suppressor (AES) [99] providing echo free half-duplex communication. If echo suppression is carried out independently at each frequency of a short-time spectral domain, a good degree of duplexity can be achieved. Recently, AES approaches have been introduced [100,101] that are similar to the residual echo postfiltering as presented in Section 4.30.2, while completely discarding the acoustic echo canceler (AEC) part. These approaches do not require the identification of the room impulse response as an FIR filter, but model parametrically the echo path with a delay and a single real-valued gain at each frequency of short-time spectra, resulting in lower computational complexity than when using a precise echo path estimation.

Conventional cancellation methods to cope with acoustic echoes have been successfully implemented, see Section 4.30.1. In practice, however, the achievable echo attenuation for these conventional approaches is not sufficient due to, e.g., the echo tail effect (modeling of too short a portion of the echo path), nonlinear echo components caused by vibration effects or the nonlinear behavior of low-cost audio hardware, and convergence problems in case of highly time varying echo paths [102]. Therefore, AEC are usually combined with a suppression scheme to remove residual echoes which the AEC lets through [35]. Commonly, the suppression of residual echoes is performed in a frequency selective way [27,35,58,62,64]. Indeed, virtually all acoustic echo cancellation systems use such a postfilter because they fail too often to reduce the echo to become sufficiently inaudible.

Author's personal copy

Such a suppressor greatly improves the echo attenuation, but contrary to the linear echo canceler alone, the resulting audio quality and double talk performance often suffer from it: Distortions on the desired signal are more likely to appear due to spectral suppression with a requirement to be aggressive enough to suppress all residual echoes [103]. Consequently, when high echo attenuation is required, the linear echo canceler can only cover a small fraction of the needed attenuation compared to the echo suppressor, and thus, the AEC is made unnecessary when the audio quality is bounded to what the suppressor yields. The benefit of purely subtractive operation by the linear echo canceler is thus restricted and an AES alone can be advantageously implemented in this case.

Now, we first review the general approach of AES as introduced in [101], before we present a complete implementation of an improved AES.

4.30.3.1 Alternative problem statement

Recall from Section 4.30.2.1 that the microphone signal $y(n)$ is composed of the near-end signal $s(n)$ and the acoustic echo signal that results from the feedback of the loudspeaker signal $x(n)$, i.e.,

$$y(n) = h_n * x(n) + s(n), \tag{30.56}$$

where h_n is the room impulse response and $*$ denotes convolution, as in (30.4). Generally, a room impulse response h_n can be decomposed into a direct sound, early reflections, and late reverberation. Here, only a global delay parameter τ and a filter g_n are used to model parametrically the echo path in order to capture direct sound and early reflections. The microphone signal $y(n)$ can thus be approximated by:

$$y(n) = g_n * x(n - \tau) + s(n). \tag{30.57}$$

As illustrated in Figure30.4, short-time discrete Fourier transform (STFT) spectra are computed from the loudspeaker and microphone signals. The STFT-domain representation of (30.57) is then given by

$$Y(m, \nu) = G(m, \nu)X_\tau(m, \nu) + S(m, \nu), \tag{30.58}$$

where m is the block time index and ν denotes the frequency index. $X_\tau(m, \nu)$ is the STFT-domain correspondence of the loudspeaker signal $x(n)$ delayed by τ samples (30.57). An estimate of the echo power spectrum can be obtained by applying an estimated delay τ and an estimate of the magnitude-squared filter $|G(m, \nu)|^2$ to the loudspeaker signal power spectrum, i.e.,

$$\left|\widehat{Y}(m, \nu)\right|^2 = \left|\widehat{G}(m, \nu)\right|^2 |X_\tau(m, \nu)|^2. \tag{30.59}$$

Since in practice the echo transfer function $|G(m, \nu)|$ is not known *a priori*, it has to be replaced by the estimate $|\widehat{G}(m, \nu)|$, corresponding to a real-valued gain at each frequency. In this model, late reflections are not estimated explicitly, but they are later considered by specific time-smoothing applied to the echo power spectrum estimate $|\widehat{Y}(m, \nu)|^2$. Then, the actual acoustic echo suppression, derived from the echo estimate (30.59), is performed by modifying the magnitude of the STFT of the microphone signal $Y(m, \nu)$, while keeping its phase unchanged. This can be expressed by

$$\widehat{S}(m, \nu) = F(m, \nu)Y(m, \nu), \tag{30.60}$$

where $F(m, \nu)$ represents a real valued, positive gain factor in each bin.

Author's personal copy

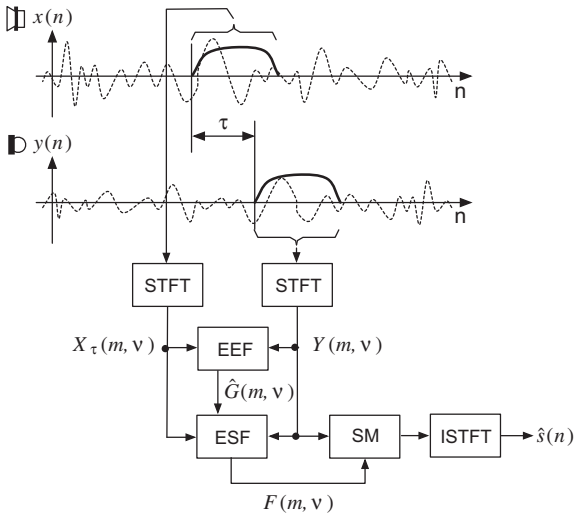


FIGURE 30.4

Basic block diagram of a short-time spectral domain acoustic echo suppressor. STFT, ISTFT, EEF, ESF, and SM stand for short-time Fourier transform, its inverse, echo estimation filter, echo suppression filter and spectral modification, respectively.

In the following, the echo path estimate in (30.59) is denoted as the *echo estimation filter* (EEF). Since the EEF is only a real-valued magnitude filter, it ought to be possible to estimate it without phase sensitivity. In [104] a technique was proposed to estimate the EEF based on power spectral fluctuations making the EEF insensitive to the phase relation between loudspeaker and microphone signals, whereas conventional echo path and EEF estimation processes are known to fail in such scenarios [105]. The suppression of echoes (30.60) is then implemented analogously to a Wiener filter [10], referred to as the *echo suppression filter* (ESF), based on a short-time power spectrum estimate of the echo. A weakness of the described AES systems is that, since a short-time ESF is applied on the microphone signal, the suppression of a low SNR echo signal tends to introduce artifacts, such as so-called "musical noise" artifacts [106]. To mitigate this problem, smoothing is applied to the echo estimate (30.59) and to the final gain filter (30.60).

From the above discussion we conclude that there are two important tasks included in acoustic echo suppression: On one hand, a suitable EEF has to be estimated in order to obtain a good estimate of the spectral components of the echo signal included in the microphone signal. On the other hand, an appropriate computation rule for the ESF is required that maximizes the echo suppression while keeping the distortions on the desired near-end signal as low as possible. Solutions to these two tasks are presented in the next two sections. The estimation of the EEF is thoroughly discussed next. Since the associated delay τ is also not known in advance, it also has to be estimated. We consider only one single delay for all frequencies. The estimation of the delay is straightforward when using correlation

Author's personal copy

methods, e.g. [107], and is not discussed further here. Then, the derivation of the final ESF is described later.

4.30.3.2 Echo path estimation

The computation of the echo estimation filter (EEF) is a crucial part of acoustic echo suppression. The estimation of the echo power spectrum $|\widehat{Y}(m,\nu)|^2$ is achieved based on the observable loudspeaker signal and an EEF (30.59), i.e., an estimate of the echo transfer function $G(m,\nu)$.

As proposed in [101], a straightforward solution for estimating $G(m,\nu)$ results from the signal model (30.58). Assuming that the near-end speaker is silent, (30.58) implies that the EEF can be estimated as the magnitude of the least squares estimator (Wiener filter),

$$\widehat{G}_{\mathrm{w}}(m,\nu) = \left| \frac{\mathcal{E}\left\{Y(m,\nu)X_\tau^*(m,\nu)\right\}}{\mathcal{E}\left\{X_\tau(m,\nu)X_\tau^*(m,\nu)\right\}} \right|, \tag{30.61}$$

where * denotes the complex conjugate operator. Since the acoustic echo path is likely to vary in time, $\widehat{G}_{\mathrm{w}}(m,\nu)$ is estimated iteratively by

$$\widehat{G}_{\mathrm{w}}(m,\nu) = \left| \frac{\Psi_{YX_\tau^*}(m,\nu)}{\Psi_{X_\tau X_\tau^*}(m,\nu)} \right|, \tag{30.62}$$

where

$$\Psi_{YX_\tau^*}(m,\nu) = \alpha Y(m,\nu)X_\tau^*(m,\nu) + (1-\alpha)\Psi_{YX_\tau^*}(m-1,\nu), \tag{30.63}$$

$$\Psi_{X_\tau X_\tau^*}(m,\nu) = \alpha X_\tau(m,\nu)X_\tau^*(m,\nu) + (1-\alpha)\Psi_{X_\tau X_\tau^*}(m-1,\nu), \tag{30.64}$$

and $\alpha \in [0,1]$ is determined as a function of the desired smoothing time constant T,

$$\alpha = 1 - \exp\left(-\frac{K}{T f_s}\right), \tag{30.65}$$

where f_s is the sampling frequency and K the STFT window hop size, $T = 1.5$ s is a reasonable value.

The above technique effectively estimates the echo path transfer function and takes the magnitude thereof to obtain the real-valued EEF. Whenever the phase changes abruptly, such as during echo path changes, this EEF estimation has to re-converge. To make (30.61) insensitive to phase variations, correlations are modified to be computed from the power spectra rather than from the complex spectra, i.e.,

$$\widehat{G}_b^2(m,\nu) = \frac{\mathcal{E}\left\{|X_\tau(m,\nu)|^2|Y(m,\nu)|^2\right\}}{\mathcal{E}\left\{|X_\tau(m,\nu)|^2|X_\tau(m,\nu)|^2\right\}}. \tag{30.66}$$

In order to illustrate that the proposed technique is insensitive to phase variations in the echo path, both EEF estimates, (30.61) and (30.66), are compared. The proposed methods are implemented using a discrete short-time Fourier transform (STFT), running at a sampling rate of 16 kHz. A 512-tap FFT is used using sine analysis and synthesis windows and successive windows have an overlap of 50%.

Author's personal copy

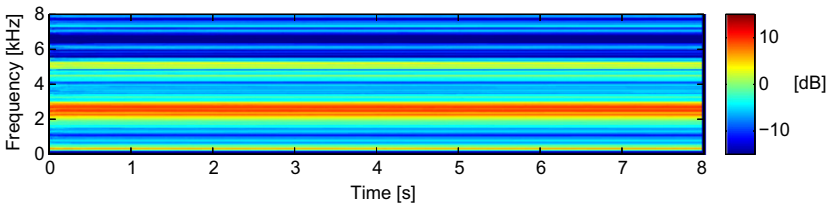


FIGURE 30.5

The true EEF $G(m,\nu)$ as it would be ideally estimated.

In order to simulate realistic scenarios, a measured room impulse response h_n of length 64 ms is used to model the echo path. Also, as reference, the true EEF $G(m, \nu)$ is shown in Figure 30.5 as it would be ideally estimated over the complete simulation time. The simulations consider a far-end signal with additive Gaussian noise with a SNR of 24 dB. The microphone signal contains the echo and near-end Gaussian noise with the same SNR. Figure 30.6 shows, for three different phase variations, the EEF computed as in (30.61) and (30.66). For Panels (a) and (d) a change in the phase of the room impulse response h_n after 4 seconds has been simulated, resulting in the EEF filters $\widehat{G}_{\mathrm{w}}(m, \nu)$ and $\widehat{G}_b(m, \nu)$, respectively. While the EEF based on complex spectra diverges from the desired filter when the phase of h_n is modified, the EEF based on power spectra stays similar to the true EEF in Figure 30.5. Also, a sampling rate mismatch of 1 Hz between loudspeaker and microphone signals was simulated. Panels (b) and (e) show the resulting EEF filters $\widehat{G}_{\mathrm{w}}(m, \nu)$ and $\widehat{G}_b(m, \nu)$, respectively, where only the second stays similar to the true EEF. The observations are the same between the EEF filters $\widehat{G}_{\mathrm{w}}(m, \nu)$ and $\widehat{G}_b(m, \nu)$, in Panels (c) and (f), where random loss of 4 samples in the loudspeaker signal was simulated. Figure 30.6 indicates that the proposed EEF estimate $\widehat{G}_b(m, \nu)$ converges quickly and is hardly affected by the phase and time modifications of the echo path. In contrast to that, the conventional EEF estimation method $\widehat{G}_{\mathrm{w}}(m, \nu)$ does not converge since it relies on phase information.

However, assuming that the non-zero near-end signal $S(m, \nu)$ and the far-end signal $X_\tau(m, \nu)$ are statistically independent and zero mean, it follows from (30.58) that $\widehat{G}_b(m, \nu)$ according to (30.66) gives

$$\widehat{G}_b^2(m, \nu) = |G(m, \nu)|^2 + \mathcal{E}\left\{|S(m, \nu)|^2\right\} \frac{\mathcal{E}\left\{|X_\tau(m, \nu)|^2\right\}}{\mathcal{E}\left\{|X_\tau(m, \nu)|^4\right\}}, \tag{30.67}$$

as demonstrated in Section 4.30.3.5. Obviously, any non-negligible near-end signal $S(m, \nu)$ included in the microphone signal $Y(m, \nu)$ leads to a positive bias in the estimate $\widehat{G}_b(m, \nu)$. The biased EEF leads to too large estimates of the echo power in the spectrum. From (30.67) it follows that this effect is especially prominent in case of high levels of the near-end signal $S(m, \nu)$. In the following, we additionally describe a method to compute the EEF based on power spectra, however, without bias.

Author's personal copy

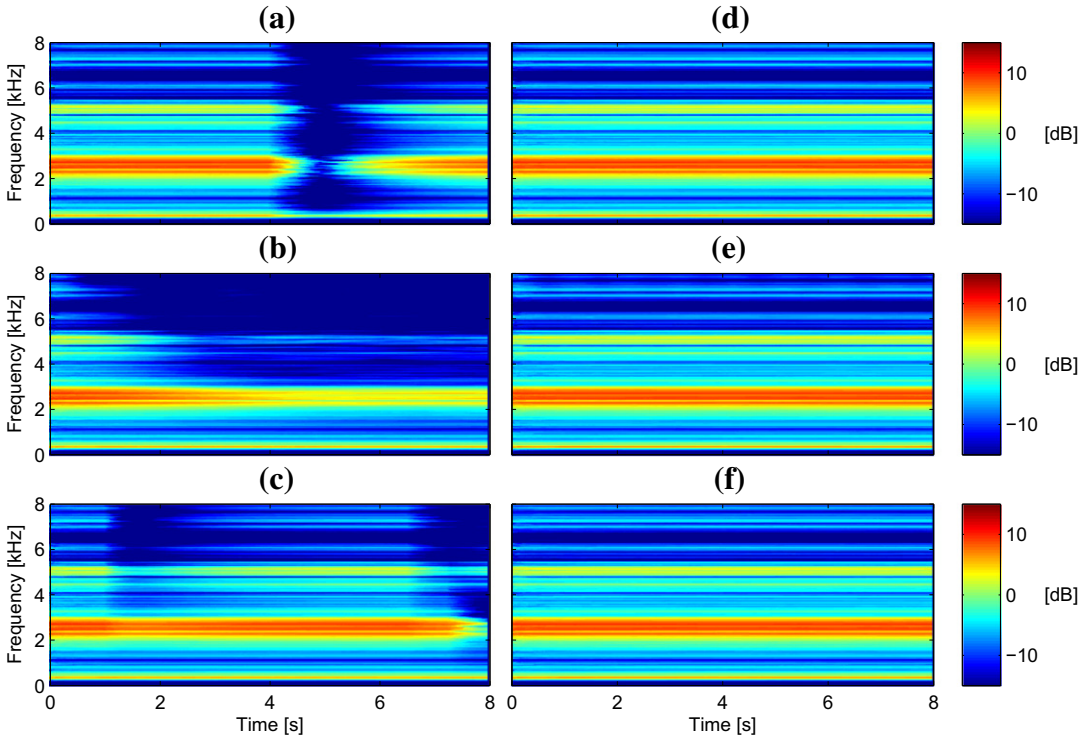


FIGURE 30.6

Effect of phase variations in the room impulse response h_n on $\widehat{G}_{\mathrm{w}}(m,\nu)$ (on the left) and $\widehat{G}_{\mathrm{b}}(m,\nu)$ (on the right). (a) and (d) Effect of a phase response modification after 4 s. (b) and (e) Effect of a sampling rate mismatch of 1 Hz between loudspeaker and microphone signals. (c) and (f) Effect of random loss of four samples in the loudspeaker signal.

This is achieved by estimating $G(m,\nu)$ based on temporal fluctuations of the power spectra computed according to

$$\widetilde{Y}(m,\nu) = |Y(m,\nu)|^2 - \mathcal{E}\left\{|Y(m,\nu)|^2\right\}, \tag{30.68}$$

$$\widetilde{X}_\tau(m,\nu) = |X_\tau(m,\nu)|^2 - \mathcal{E}\left\{|X_\tau(m,\nu)|^2\right\}. \tag{30.69}$$

In practice, the expectation operator $\mathcal{E}\{.\}$ is implemented as single pole temporal averaging analogously to (30.63). Note that the smoothing time constant used in (30.68) and (30.69) is chosen smaller than the time constant T in (30.65), i.e., about a few hundred of milliseconds. The estimation of the EEF is then performed analogously to (30.66), but based on the fluctuating spectra of the loudspeaker and

Author's personal copy

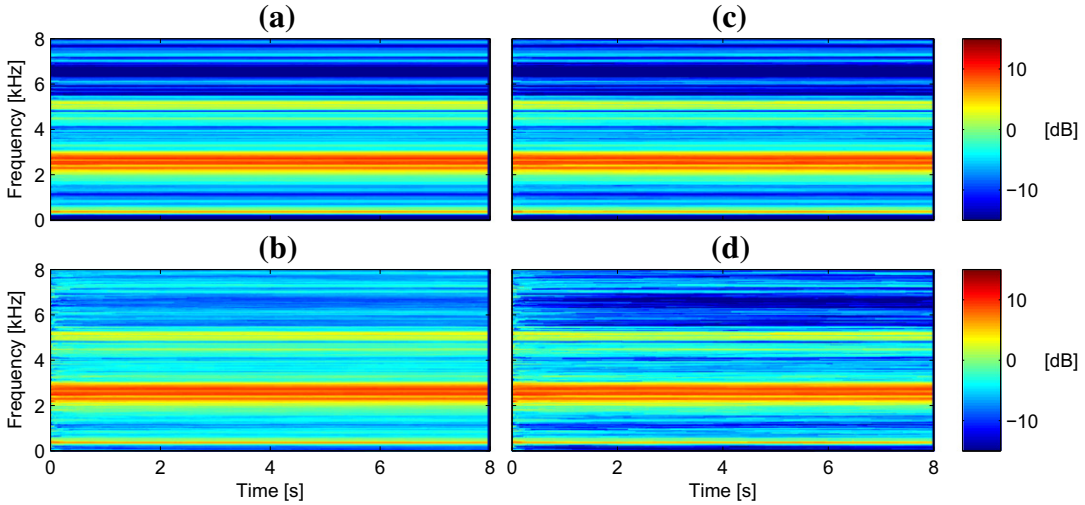


FIGURE 30.7

Biased (on the left) and unbiased (on the right) EEF estimates for 24 dB and 6 dB SNR in Panels (a), (d) and (c), (d), respectively.

microphone signals:

$$\widehat{G}^2(m,\nu) = \frac{\mathcal{E}\left\{\widetilde{X}_\tau(m,\nu)\widetilde{Y}(m,\nu)\right\}}{\mathcal{E}\left\{\widetilde{X}_\tau(m,\nu)\widetilde{X}_\tau(m,\nu)\right\}}. \tag{30.70}$$

This, as demonstrated in Section 4.30.3.5, yields an un-biased echo estimation filter,

$$\widehat{G}^2(m,\nu) = |G(m,\nu)|^2. \tag{30.71}$$

It is important to note that the fluctuating power spectra are only used for the estimation of $G(m,\nu)$. The computation of the echo suppression filter $F(m,\nu)$, in (30.60), is still based on the original power spectra of the loudspeaker and microphone signals. Based on the simulations as defined previously, we show the bias resulting from the EEF estimate using power spectra (30.66). A far-end signal with a SNR of 24 dB is considered, while the microphone signal contains the echo and near-end Gaussian noise with two different SNRs: 24 dB and 6 dB. Figure 30.7 shows in the two left panels the biased EEF estimates $\widehat{G}_b(m,\nu)$, and in the two right panels, the unbiased EEF estimates $\widehat{G}(m,\nu)$, for 24 dB and 6 dB SNR, respectively. While the unbiased EEF estimates $\widehat{G}(m,\nu)$ are similar for all SNR conditions, the biased EEF estimates $\widehat{G}_b(m,\nu)$ are more impaired the lower the SNR is.

Eventually, in order to prevent the EEF (30.71) from diverging, when near-end speech is active, a two echo path model [43] is used. One background path comprising a fully adaptive echo estimation filter, and one foreground path which comprises the echo estimation filter effectively used for computation of the echo power spectrum. The values of the foreground echo estimation filter are refreshed by those of the background one based on the performance of the algorithm. Also two voice activity detectors

Author's personal copy

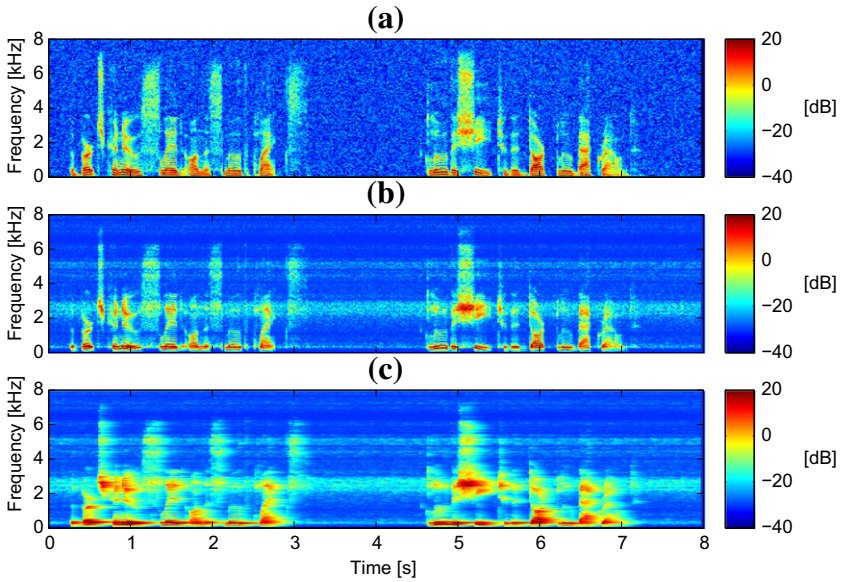


FIGURE 30.8

A far-end speech signal is considered in Panel (a). The resulting echo estimates corresponding to (30.59) and (30.72) are respectively shown in Panels (b) and (c). The used exponential decay α_{RT} corresponds to a time constant of 60 ms.

(VADs) at far-end and near-end sides [2], respectively, are used to discriminate whenever far-end and/or near-end speech is active.

4.30.3.3 Echo suppression filter

The estimate of the echo spectrum according to (30.59) covers only a fraction of the length of the true echo path corresponding to direct sound and early reflections. To cope with the echo components resulting from late reverberation, a temporal smoothing is applied to the echo spectrum estimate in order to mimic typical exponential decay of late reflexions. This is achieved by applying recursively a forgetting factor on the echo power spectrum estimate of previous frame

$$|\widehat{Y}(m,\nu)| = \max\left\{|\widehat{Y}(m,\nu)|, \alpha_{\mathrm{RT}}|\widehat{Y}(m-1,\nu)|\right\}, \tag{30.72}$$

where $\alpha_{\mathrm{RT}} \in [0,1]$ is determined as a function of the amount of late reverberation and is computed similarly to (30.65). Based on the simulations as defined previously, and considering a far-end speech signal $X(m,\nu)$ shown in Panel (a) of Figure 30.8, the corresponding echo estimate (30.59) is shown in Panel (b), while the smoothed estimate (30.72) is shown in Panel (c). The described temporal smoothing does not change signal dynamic but models the exponential decay of the echo.

Author's personal copy

From the echo estimate (30.72), the optimum values for the echo suppression filter (ESF) $F(m, \nu)$ can be derived by minimizing the contribution of the echo components $|\widehat{Y}(m, \nu)|^2$ to the output signal $\widehat{S}(m, \nu)$ in the mean square error (MSE) sense. Since the near-end signal $S(m, \nu)$ and the loudspeaker signal $X(m, \nu)$ were assumed to be statistically independent, we obtain [10]

$$F_{\text{opt}}(m, \nu) = \frac{\mathcal{E}\left\{|Y(m, \nu)|^2\right\} - \mathcal{E}\left\{|\widehat{Y}(m, \nu)|^2\right\}}{\mathcal{E}\left\{|Y(m, \nu)|^2\right\}}. \tag{30.73}$$

A practical approach for the computation of the ESF is based on generalized, instantaneous versions of (30.73). In [101] it has been proposed to use the power spectral subtraction approach analogously to [108]:

$$F(m, \nu) = \frac{|Y(m, \nu)|^2 - \beta|\widehat{Y}(m, \nu)|^2}{|Y(m, \nu)|^2}, \tag{30.74}$$

where β represents a design parameter to control the amount of echo to be suppressed [109]. $F(m, \nu)$ can be considered as an estimate of $F_{\text{opt}}(m, \nu)$ according to (30.73). To prevent residual echoes, the ESF in (30.74) is computed to attenuate the microphone signal aggressively such that no residual echo remains. This is, e.g., achieved by intentionally over-estimating the echo power spectrum (by choosing $\beta > 1$), but also by applying a suitable time-smoothing to $F(m, \nu)$. These design parameters have an important role to address attenuation of residual echoes, resulting from long echo paths, non-linearities, etc. Considering a near-end speech signal as shown in Figure 30.9, the computation of ESF is simulated with the same far-end signal as in Panel (a) of Figure 30.8 and the echo path as defined in Figure 30.5. The resulting ESF is shown in Panel (b) of Figure 30.9, where the contribution for the echo is removed and the time-frequency tiles corresponding to near-end signal are kept untouched as shown by Panel (c).

The acoustic echo suppression method, as presented above, is derived analogously to spectral subtraction used for stationary noise suppression [106]. And since echo control applications often require in the same time noise suppression, both echo and noise suppression can be advantageously combined to minimize the resulting distortions on the processed signal like "musical noise" artifacts [110].

4.30.3.4 Perceptual acoustic echo suppression

In order to further reduce computational complexity, the AES processing is not carried out on each STFT frequency bin separately. The uniformly spaced spectral coefficients can be grouped into a number of non-overlapping partitions, similar as in [111]. Each such partition (group of bins) corresponds to one subband in which the processing is carried out. The bandwidth of each group of bins is chosen such that it roughly follows the frequency resolution of the human auditory system. The partition bandwidth for example is chosen to be approximately two times the equivalent rectangular bandwidth (ERB) [112], resulting, for example, in 16 partitions for 16 kHz as shown in Panel (a) of Figure 30.10. The different statistics and filters are then only computed once for each partition, instead of once for each STFT frequency bin, resulting in lower computational complexity. Prior to applying the partition echo suppression filter (ESF) to the uniform signal STFT spectrum, it has to be interpolated. An example of this interpolation is illustrated in Panel (b) of Figure 30.10: At one frame, the values for partitions are indicated as points (•) and the resulting interpolated values, obtained by interpolation, are indicated as line. In this way, the proposed interpolation smoothes the filter values over frequency reducing artifacts

Author's personal copy

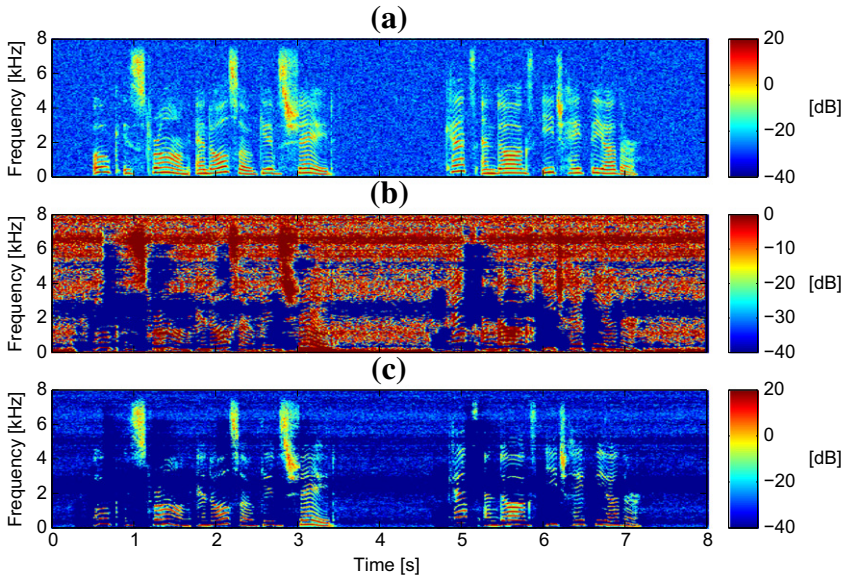


FIGURE 30.9

The desired near-end speech signal is shown in Panel (a). The ESF is shown in Panel (b) with a far-end signal chosen to be the same as in Panel (a) of Figure 30.8. The resulting output signal $\widehat{S}(m,\nu)$ is shown in Panel (c)

which would result from high fluctuations of the filter. This approach, based on frequency resolution of the human auditory system, is referred as to perceptual acoustic echo suppression (PAES) [113].

4.30.3.5 Derivation of echo estimation filters

The present subsection gives the derivation which lead to the biased and unbiased echo estimation filters (EEF), (30.67) and (30.71), respectively. First, we consider a general result holding for statistically independent processes. Let A and B be two independent statistical random processes, and f and g two arbitrary functions. Then,

$$\mathcal{E}\{f(A)g(B)\} = \mathcal{E}\{f(A)\}\,\mathcal{E}\{g(B)\} \tag{30.75}$$

holds [114].

Regarding the EEF derivation itself, it is reasonable to assume that loudspeaker signal $X_\tau(m,\nu)$ and near-end signal $S(m,\nu)$ are statistically independent, zero-mean random processes. In the following, the indexes m and ν are discarded for presentational simplicity. From (30.58) it follows that the power

Author's personal copy

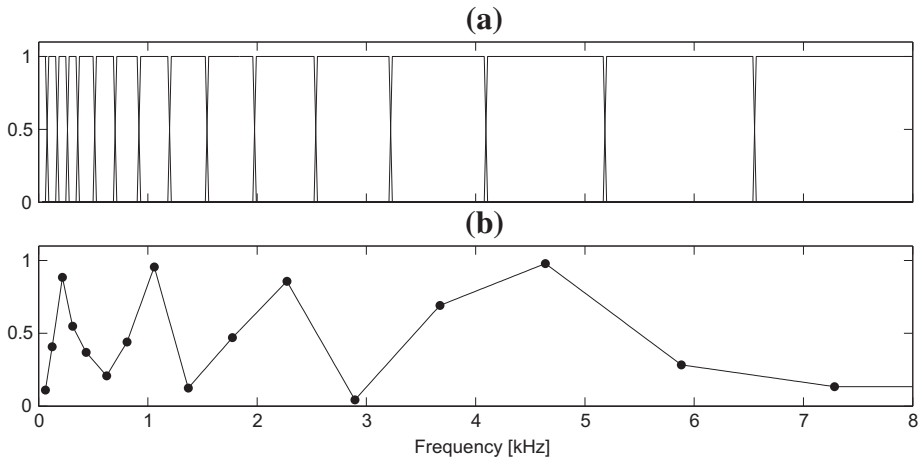


FIGURE 30.10

Panel (a) shows how the STFT frequency bins are grouped to obtain partitions mimicking the frequency resolution of the human auditory system. Panel (b) shows an example of interpolation of the ESF $F(m,\nu)$ from partitions to STFT frequency bins to be applied on the microphone spectrum $Y(m,\nu)$.

spectrum $|Y|^2$ can be written as:

$$\begin{aligned} |Y|^2 &= YY^* \\ &= (GX_\tau + S)(GX_\tau + S)^* \\ &= |G|^2|X_\tau|^2 + |S|^2 + G^*X_\tau^*S + GX_\tau S^*. \end{aligned} \tag{30.76}$$

Since both, X_τ, and S are statistically independent, zero-mean processes,

$$\mathcal{E}\left\{|Y|^2\right\} = |G|^2\mathcal{E}\left\{|X_\tau|^2\right\} + \mathcal{E}\left\{|S|^2\right\}. \tag{30.77}$$

The EEF is estimated by (30.66), whose numerator is

$$\mathcal{E}\left\{|Y|^2|X_\tau|^2\right\} = \mathcal{E}\left\{\left(|G|^2|X_\tau|^2 + |S|^2 + G^*X_\tau^*S + GX_\tau S^*\right)|X_\tau|^2\right\}, \tag{30.78}$$

considering 30.75, it leads to

$$\mathcal{E}\left\{|Y|^2|X_\tau|^2\right\} = |G|^2\mathcal{E}\left\{|X_\tau|^4\right\} + \mathcal{E}\left\{|S|^2\right\}\mathcal{E}\left\{|X_\tau|^2\right\}. \tag{30.79}$$

Furthermore, the denominator of (30.66) is

$$\mathcal{E}\left\{|X_\tau|^2|X_\tau|^2\right\} = \mathcal{E}\left\{|X_\tau|^4\right\}. \tag{30.80}$$

Author's personal copy

The EEF according to (30.79) and (30.80), thus, yields

$$\widehat{G}_b^2 = |G|^2 + E|S|^2 \frac{\mathcal{E}\left\{|X_\tau|^2\right\}}{\mathcal{E}\left\{|X_\tau|^4\right\}}. \tag{30.81}$$

As can be seen, the near-end signal S introduces a bias term into the estimate of the EEF. Furthermore, (30.81) implies that the bias in the EEF increases with increasing near-end signal variance.

In the proposed method, in order to avoid the bias introduced in (30.66), the EEF in (30.70) is computed based on the temporal fluctuations of the power spectra (30.68) and (30.69). The numerator of the EEF (30.70) is given by

$$\begin{aligned}\mathcal{E}\left\{\widetilde{Y}\widetilde{X}_\tau\right\} = \mathrm{E}\Big\{\Big(&|G|^2|X_\tau|^2 + |S|^2 + G^* X_\tau^* S + G X_\tau S^* \\ &- |G|^2\mathcal{E}\left\{|X_\tau|^2\right\} - \mathcal{E}\left\{|S|^2\right\}\Big)\left(|X_\tau|^2 - \mathcal{E}\left\{|X_\tau|^2\right\}\right)\Big\}.\end{aligned} \tag{30.82}$$

This simplifies to

$$\mathcal{E}\left\{\widetilde{Y}\widetilde{X}_\tau\right\} = \mathcal{E}\left\{|G|^2|X_\tau|^4 - 2|G|^2|X_\tau|^2\mathcal{E}\left\{|X_\tau|^2\right\} + |G|^2\mathcal{E}\left\{|X_\tau|^2\right\}^2\right\}, \tag{30.83}$$

and finally, considering (30.75),

$$\mathcal{E}\left\{\widetilde{Y}\widetilde{X}_\tau\right\} = |G|^2\left(\mathcal{E}\left\{|X_\tau|^4\right\} - \mathcal{E}\left\{|X_\tau|^2\right\}^2\right). \tag{30.84}$$

Also, the denominator of (30.70) is

$$\begin{aligned}\mathcal{E}\left\{\widetilde{X}_\tau\widetilde{X}_\tau\right\} &= \mathcal{E}\left\{\left(|X_\tau|^2 - \mathcal{E}\left\{|X_\tau|^2\right\}\right)\left(|X_\tau|^2 - \mathcal{E}\left\{|X_\tau|^2\right\}\right)\right\} \\ &= \mathcal{E}\left\{|X_\tau|^4\right\} - \mathcal{E}\left\{|X_\tau|^2\right\}^2.\end{aligned} \tag{30.85}$$

Thus, the EEF according to (30.70) yields

$$\widehat{G}^2 = |G|^2. \tag{30.86}$$

Note that (30.70) leads to an unbiased estimate of the echo power transfer function also in case of near-end signal included in the microphone signal.

4.30.4 Multichannel acoustic echo cancellation

Multichannel acoustic echo cancellation (MCAEC) is a key technology whenever hands-free and full-duplex communication in modern systems with multichannel sound reproduction is required. For various applications, such as home entertainment, virtual reality (e.g., games, simulations, training), or advanced teleconferencing, multimedia terminals with an increased number of audio channels for sound reproduction are highly desirable (e.g., stereo, 5.1 surround systems, or even beyond). Although the basic

Author's personal copy

principle of echo cancellation has been well known for several decades, the multichannel case poses some additional and fundamentally different challenges. Moreover, there are even some notable differences between the two-channel case and the general multichannel case which has been addressed bit by bit only in recent years. The aim of this section is twofold. On the one hand, after a brief review of the problem of multichannel acoustic echo cancellation, this section gives an outline of how the problem may be tackled based on some fundamental principles. In this sense, the presentation in this section brings together ideas from the theory on signals and systems, information theory, psychoacoustics, and also wave physics. Based on this framework, and as the other main contribution, we present in this section some recent advances in the field of MCAEC. Thereby, important issues in the case of more than two channels are emphasized. Finally, as an outlook, we touch on some ongoing work towards MCAEC for massive multichannel sound reproduction, such as wave field synthesis.

4.30.4.1 Description of signals and systems

Acoustic echo cancellation has already been discussed extensively for stereo sound reproduction (e.g., [115–118]). Only in recent years, AEC has been realized for more than two reproduction channels [26, 119, 120]. Figure 30.11 describes a typical scenario for stereo and multi-channel AEC. In a transmission room, a sound source (e.g., a speaker) is picked up by P microphones ($P = 2$ for stereo). The microphone signals are transmitted to a receiving room and reproduced via P loudspeakers. At the same time, a microphone in the receiving room picks up speech from a local user. In order to avoid an echo of the loudspeaker signals $x_i(n)$ in the transmission room, the AEC attempts to cancel the additional contributions by the loudspeakers to the microphone by subtracting filtered versions of the loudspeaker signals from the microphone signal. This generally requires that cancellation filters (assumed to be

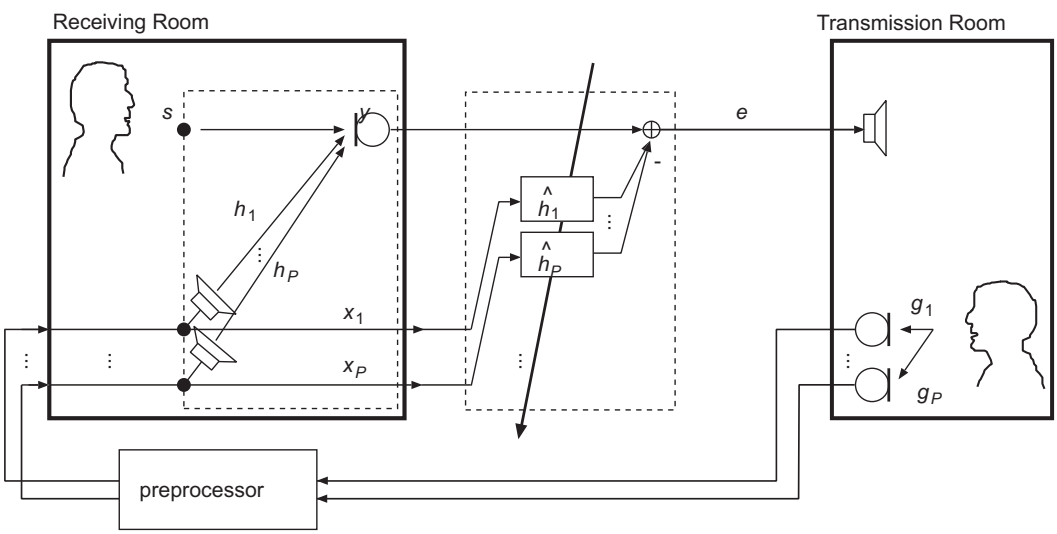


FIGURE 30.11

Scenario for multi-channel AEC.

Author's personal copy

length-N FIR filters) are dynamically adjusted by an adaptation algorithm to achieve minimum error signal $e(n)$ and thus optimum cancellation. This is the case when the adaptive cancellation filters

$$\hat{\mathbf{h}}_i(n) = \left[\hat{h}_{i,1}(n), \ldots, \hat{h}_{i,N}(n)\right]^T, \quad i = 1, 2, \ldots, P \tag{30.87}$$

accurately model the impulse responses $\mathbf{h}_i$ from the emitting speakers to the microphone.

It has been shown for stereo AEC that a so-called *non-uniqueness problem* exists [121]: If the loudspeaker signals are strongly correlated, then the adaptive filters generally converge to a solution that does not correctly model the transfer functions between the speakers and the microphone, but merely optimizes echo cancellation for the given particular loudspeaker signals [28]. This is due to the fact that the observation model $y = h_1 * x_1 + \cdots + h_P * x_P$ does not provide us with enough linearly independent equations for resolving the unknowns h_i and the problem is thus underdetermined. As a consequence, a change in the characteristics of the loudspeaker signals (e.g., due to a change of the geometric position of the sound source in the transmission room) will result in a breakdown of the echo cancellation performance and requires new adaptation of the cancellation filters.

From a statistical point of view, the high cross-correlations between the loudspeaker signals lead to a highly ill-conditioned tap-input correlation matrix $\mathbf{R}_{\mathbf{xx}}(n)$ in the normal equation, $\mathbf{R}_{\mathbf{xx}}(n)\mathbf{h}(n) = \mathbf{r}_{\mathbf{xx}}(n)$, to be solved for the minimization of $\mathcal{E}\{e^2(n)\}$ [115], where

$$\begin{aligned}
\mathbf{R}_{\mathbf{xx}}(n) &= \mathcal{E}\left\{\mathbf{x}(n)\mathbf{x}^T(n)\right\} \qquad (30.88)\\
&= \begin{bmatrix} \mathbf{R}_{\mathbf{x}_1\mathbf{x}_1}(n) & \cdots & \mathbf{R}_{\mathbf{x}_1\mathbf{x}_P}(n) \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{\mathbf{x}_P\mathbf{x}_1}(n) & \cdots & \mathbf{R}_{\mathbf{x}_P\mathbf{x}_P}(n) \end{bmatrix}, \\
\mathbf{x}(n) &= \left[\mathbf{x}_1(n), \ldots, \mathbf{x}_P(n)\right], \\
\mathbf{x}_i(n) &= \left[x_i(n), \ldots, x_i(n-N+1)\right]^T, \\
\mathbf{h}(n) &= \left[\mathbf{h}_1(n), \ldots, \mathbf{h}_P(n)\right]. \qquad (30.89)
\end{aligned}$$

To tackle this challenging problem of ill conditioning, various techniques have been proposed mainly in the stereo context so far. They can be distinguished into two different classes representing separate system components as shown, e.g., in [116]:

(a) Application of a robust and fast converging adaptation algorithm taking all cross-correlations into account.

(b) Preprocessing of the signals transmitted from the transmission room prior to their reproduction in the receiving room in order to *partially decorrelate* all channels relative to each other.

We then face the two conflicting requirements that, on the one hand, the preprocessing must not introduce any objectionable artifacts into the reproduced audio signals while, on the other hand, we require a decorrelation for convergence enhancement. Therefore, a systematic design for MCAEC based on first principles of coefficient estimation and optimization, together with a complete stochastic signal description, and considering human auditory perception, is necessary. The structure of this section is motivated by a step-by-step incorporation of these principles. Within this framework, we place recent advances in MCAEC with emphasis on more than two reproduction channels, and deduce various new insights and practical results.

Author's personal copy

4.30.4.2 Elements from estimation and information theory

In general, to optimally exploit the information contained in the involved signals, the coefficient estimation process should take into account all their fundamental stochastic properties: *Nongaussianity, nonwhiteness, nonstationarity*. A suitable broadband signal formulation for this purpose was developed within the so-called TRINICON framework for adaptive multiple-input and multiple-output (MIMO) filtering [55,57,122,123], as already mentioned in Section 4.30.1.

In [56], the AEC problem was linked explicitly to the more general *MIMO system identification and signal separation problem* as addressed by TRINICON, and as illustrated by the two dashed boxes in Figure 30.11. The left and right dashed boxes correspond to a MIMO mixing system and a corresponding MIMO demixing system, respectively. The demixing system follows rigorously from the ideal MIMO separation solution derived in [124,125]. This formal connection facilitates the introduction of stochastic signal models in the form of multivariate probability densities which capture the temporal structure by multiple time lags and the nonstationarity by time-varying correlation matrices.

The TRINICON optimization criterion for the case of separation (and system identification) problems is based on minimizing the information-theoretic quantity of *mutual information* between the output channels of the demixing MIMO system using the multivariate densities mentioned above. In the special case of AEC, we separate the contributions of the loudspeaker signals from the error signal $e(n)$ at the AEC output (Figure 30.11 and [56]). In the special case of Gaussian signals, this separation process corresponds to a simultaneous block-diagonalization of the output correlation matrix for multiple time instants since the local speech $s(n)$ is assumed to be uncorrelated from the loudspeaker signals [56,123].

Following [126], we show here how to generalize the information-theoretic separation approach in [56] to *multichannel* AEC with typically highly correlated loudspeaker signals. Specifically, the output channels $x_1(n), \ldots, x_P(n)$ of the mixing system in Figure 30.11 do *not* require separation from each other. Figure30.12 illustrates the output correlation matrix $\mathbf{R}_{\mathbf{yx}}$ after mixing (left) and the corresponding *desired* structure $\mathbf{R}_{\mathbf{ex}}$ after demixing (right) for the special case of Gaussian signals and $P = 2$. Hence, the approach in [56] generalizes straightforwardly to the MCAEC case by just using this modified matrix partitioning.

In [56] the update equations for TRINICON-based coefficient adaptation in AEC have been presented for the simple case of gradient-based optimization. However, it is known that gradient-descent algorithms (e.g., LMS/NLMS [10]) generally exhibit very slow convergence for highly correlated input signals such as in the multichannel case.

The so-called Newton-Raphson-type optimization procedure is known as the canonical method for more challenging optimization problems. As detailed in [55], a TRINICON-based Newton update can be derived in a way analogous to [52]. The Newton algorithm contains virtually all of the well-known adaptation schemes as special cases, most notably the recursive least-squares (RLS) algorithm. The important feature of Newton-type/RLS-type algorithms is that they explicitly take all input correlations (30.88) into account within their Hessian matrix [10,52] which makes them very attractive for the MCAEC application [116].

In addition to this desirable property of RLS-type algorithms, the more general TRINICON-based approach inherently leads to a *multivariate* error *nonlinearity* to take both the nongaussianity and the nonwhiteness of the near-end signal into account [56]. This provides an inherent double-talk handling and a link to the powerful concept of robust statistics, e.g., [51,53]. Moreover, the block online adaptation and block averaging obtained in [56] further speeds up the convergence (especially in MCAEC).

Author's personal copy

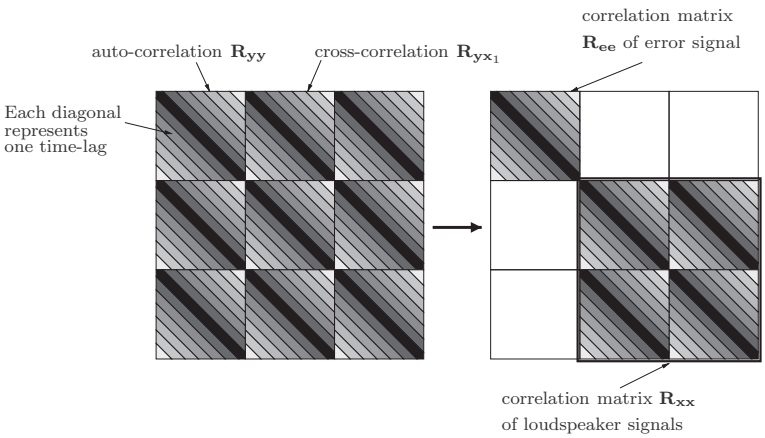


FIGURE 30.12

AEC process for second-order statistics and $P = 2$.

Note also that the general TRINICON-based approach also leads to important insights in the case of AEC for multiple microphone channels in the receiving room, as explained further in Section 4.30.4.4.

Finally, another aspect in the design of a real-time solution to MCAEC is its computational complexity. Unfortunately, straightforward implementations of RLS-type algorithms are computationally very expensive due to the required (implicit or explicit) inversion of the correlation matrix. A very efficient practical solution to this problem is to formulate the above-mentioned broadband algorithm in a mathematically rigorous way in the frequency domain, as shown, e.g., in [26,52,123], followed by the introduction of *carefully selected* approximations. The most important features of this concept of *frequency-domain adaptive filtering* (FDAF) is that in addition to the efficient use of the FFT (gains for both, adaptation and filtering), all the sub-matrices of the input correlation matrix (30.88) are approximately diagonalized by the DFT. In this way, it is possible to efficiently take into account all cross-correlations [26]. This is possible for both, second-order and higher-order statistics. A first MCAEC system for 5-channel surround sound applications, based on the multichannel FDAF algorithm has been presented in [26,119]. This real-time implementation also utilizes the concept of robust statistics [52].

4.30.4.3 Elements from psychoacoustics

As mentioned in Section 4.30.4.1, among the key requirements for the techniques to preprocess the signals transmitted from the transmission room prior to their reproduction in the receiving room is the subjective sound quality. While several of the known preprocessing techniques provide enough decorrelation to achieve proper AEC convergence in the stereo case, considerations of sound quality have frequently not been addressed adequately. In this section we first give a brief overview of the known two-channel preprocessing approaches. We then describe a recently introduced novel approach

Author's personal copy

[120], based on perceptual considerations. It easily generalizes to the multi-channel case and has been demonstrated to be effective in surround sound echo cancellation.

4.30.4.3.1 Known two-channel preprocessing approaches

A first simple preprocessing method for stereo AEC was proposed by Benesty et al. [115,127] and achieves signal decorrelation by adding non-linear distortions to the signals. While this approach features extremely low complexity, the introduced distortion products can become quite audible and objectionable, especially for high-quality applications using music signals. Moreover, the generalization of this approach to an arbitrary number of channels is not straightforward.

A second well-known approach consists of adding uncorrelated noise to the signals. In [128], this is achieved by perceptual audio coding/decoding of the signal which introduces uncorrelated quantization distortion that is masked due to the noise shaping according to the encoder's psychoacoustic model. The use of an explicit psychoacoustic model plus analysis/synthesis filterbanks is able to prevent audible distortions in audio signals and may be easily generalized to more than two channels. However, the associated implementation complexity and the introduced delay render this approach unattractive for most applications.

Other approaches employ switched/time-varying time-delays [118] or variable all-pass filtering [129] to produce a time-varying phase shift / signal delay between the two channels of a stereo AEC and thus "decorrelate" both signals. Specifically, [118] describes a preprocessing system in which the output signal switches between the original signal and a time-delayed/filtered version of it. As a disadvantage, this switching process may introduce unintended artifacts into the audio signal. Ali [129] describes a system in which an allpass preprocessor is randomly modulating its allpass filter parameter. In [130], it was proposed to apply this allpass preprocessor only to the low frequency range up to 1 kHz due to convergence requirements.

4.30.4.3.2 Psychoacoustically motivated method for the multichannel case

In order to obtain a preprocessing method offering both good decorrelation properties for the enhancement of AEC convergence *and* minimal alteration of the perceived stereo image, the method proposed in [120] is based on several considerations. From the previously discussed approaches, the time-varying modulation of the phase of the audio signal, as proposed in [118,129], is an effective method which is generally unobtrusive in its perceptual effects on audio signals as compared to other methods while avoiding computationally expensive masking models. Nonetheless, it is difficult to achieve maximum decorrelation while guaranteeing that introducing a time/phase difference between left and right channels does not result in an alteration of the perceived stereo image. Several aspects must be accounted for:

- Interaural phase/time difference is a relevant perceptual parameter for subjective perception of a sound stage [131] and has been used extensively in synthesis of stereo images (e.g., [111]). Consequently, a change in the perceived stereo image can only be avoided if the introduced time/phase difference stays below the threshold of perception, as it applies to audio signals that are reproduced via loudspeakers.
- Optimal AEC convergence enhancement can be achieved if the preprocessing introduces time/phase differences just at the threshold of perception, i.e., applies the full amount of tolerable modification.

Author's personal copy

- Known from psychoacoustics, the human sensitivity to phase differences is high at low frequencies, and gradually reduces for increasing frequencies, until it fully vanishes for frequencies above ca. 4 kHz.
- Neither a simple time delay modulation nor a low-order time-varying allpass filtering approach offers the flexibility to tailor the amount of time/phase shifting as a function of frequency, such that the full potential of perceptually tolerable change is exploited.

Hence, in contrast to earlier phase modulation approaches, the method in [120] is designed to allow a perceptually motivated frequency-selective choice of phase modulation parameters (modulation frequency, modulation amplitude, and modulation waveform) by employing analysis/synthesis filterbanks. The input audio signal is decomposed into subband signals by means of an analysis filterbank. Then, the subband phases are modified based on a set of frequency-dependent modulating signals. According to the above considerations, subbands belonging to the low frequency part of an audio signal should be left largely untouched, while subbands corresponding to frequencies above 4 kHz may be modulated heavily. The frequency-selective phase modulation amplitude can be optimized by a listening procedure. Finally, the modified spectral coefficients are converted back into a time-domain representation by a synthesis filterbank. To allow easy access to the signal's phase, a complex-valued filterbank [132] is used, and a phase modification is implemented by a complex multiplication of the subband coefficient with $e^{j\varphi(t,\nu)}$ where $\varphi(t, \nu)$ denotes the intended time-varying phase shift in subband ν. It is preferable to choose a smooth modulating function $\varphi(t, \nu)$, such as a sine wave at a relatively low frequency. Moreover, to account for the symmetry of typical multi-channel speaker setups, such as 5.1 or 7.1, the modulation of channel pairs is carried out in a complex conjugate fashion. The modulation frequencies for pairs are chosen such that they provide "orthogonal" modulation activity.

Figure30.13 shows a summary of the results of a standardized subjective MUSHRA ("MUlti Stimulus test with Hidden Reference and Anchor") listening test carried out with 10 experienced listeners in a typical surround sound listening setup. The sound quality was quantified on a scale from 0 to 100 for 5 critical music excerpts and one speech excerpt (see [120] for further details). The different

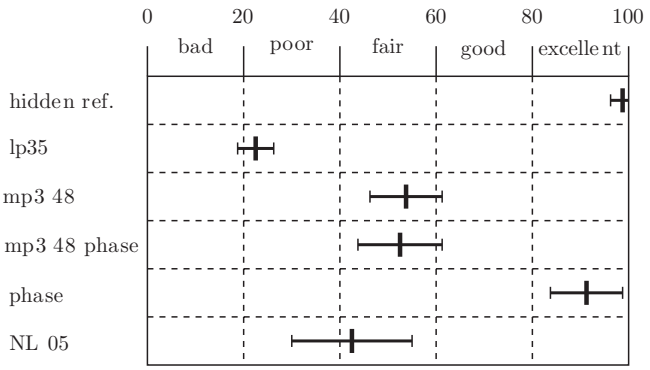


FIGURE 30.13

MUSHRA listening test results (averages and 95% confidence).

Author's personal copy

preprocessing types are the original reference and a 3.5 kHz band-limited version thereof, both required by MUSHRA, the individual channel mp3 en/decoding at 48 kbit/s ("mp3 48"), the described perceptual phase modulation method ("phase"), a combination of mp3 encoding/decoding and phase modulation ("mp3 48 phase"), and the conventional non-linear processing ("NL 05" after [115,127]). It is visible from the graph that the phase modulation method emerges as the clear winner in terms of sound quality. Note that the latter four methods were tuned for comparable convergence speeds of the adaptive algorithms.

4.30.4.4 MIMO processing and elements from wave physics

4.30.4.4.1 MIMO case for multiple microphones

So far in this section, we have focused on the case of multiple reproduction channels but only one microphone in the receiving room. The more general case of a full MIMO loudspeaker-room-microphone system appears when combining MCAEC with a microphone array, e.g., [26]. Traditionally, in this case several parallel multiple-input and single-output (MISO) systems are independently applied, which has been shown to be optimal in terms of least-squares-based coefficient estimation.

As explained in Section 4.30.4.2, TRINICON-based AEC is generally able to exploit the nonwhiteness of the signals in the receiving room (upper left sub-matrix in Figure 30.12). By further generalizing the TRINICON-based AEC to the case of MIMO loudspeaker-room-microphone systems, it is also able to exploit the *spatial* nonwhiteness in the receiving room by simultaneously taking into account all microphone signals for the adaptation process. In other words, the performance may be improved with multiple microphones.

4.30.4.4.2 Massive multichannel systems and wave physics

Current loudspeaker setups, such as the 5.1 format, still rely on a restrained listening area ("sweet spot"). A high-quality volume solution for a large listening space is offered by the wave field synthesis (WFS) method which is based on wave physics [133]. The so-called *Kirchhoff-Helmholtz integrals* which can be derived from the acoustic wave equation state that at any point within a source-free listening area, the sound pressure field can be calculated if both the sound pressure and its gradient are known on the *contour* enclosing this area. Thus, in WFS, closely spaced arrays of a large number P of individually driven loudspeakers generate a pre-specified sound field. P may lie between 20 and several hundred. An analogous approach is possible for wave field analysis (WFA) using microphone arrays.

Building a full-duplex system with this massive multichannel setup for unrestricted audio content might be considered as the supreme discipline of MCAEC research since in this case even the $P \times P$ frequency bin-wise correlation matrices of the loudspeaker driving signals are generally still large and ill-conditioned after the approximate blockwise diagonalization of (30.88) within the frequency-domain adaptive filtering (FDAF) coefficient update (cf. Section 4.30.4.2).

The basic idea of *wave-domain adaptive filtering* (*WDAF*), e.g., [134,135], is to replace the point-to-point MIMO system model by a more detailed spatial consideration exploiting wave-physics foundations as in WFS/WFA. In particular, WDAF extends the conventional FDAF approach by a suitable *spatio-temporal* transform for efficiency. Figure30.14 illustrates this two-step transformation approach from the RLS via FDAF towards WDAF in terms of the loudspeaker correlation matrix and its approximate temporal and spatio-temporal diagonalization. It can be seen that each transformation step is supposed

Author's personal copy

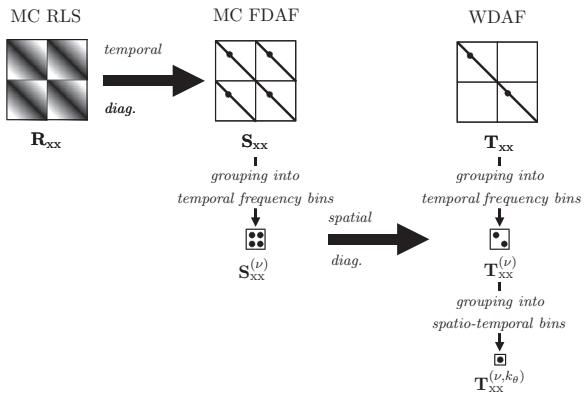


FIGURE 30.14

WDAF concept and relationship with conventional algorithms.

to achieve yet more sparseness of the multichannel correlation matrix. More sparseness then provides a more tangible basis for explicit or implicit matrix inversion.

Requirements for the spatio-temporal basis functions are that they should be orthogonal and must fulfill the acoustic wave equation (e.g., circular harmonics). Moreover, since the transducers are only placed on the contour enclosing the listening area, corresponding transformations taking into account the Kirchhoff-Helmholtz Integrals are necessary. These transformations depend on the array geometries, and for certain setups, e.g., circular arrays [134,135], they can in fact be formulated in a compact form. A rigorous formulation of RLS-type MIMO algorithms in spatio-temporal transform domains using arbitrary orthogonal bases was developed in [136] as a systematic extension of the FDAF formalism, e.g., [26,52,123].

Advantages of the approximate MIMO decoupling due to the spatio-temporal transformation are both an improved convergence and a significant complexity reduction, as shown, e.g., in [134,135]. Note also that the WDAF concept can be well applied to the general TRINICON approach. Since all microphone signals are jointly taken into account by the spatio-temporal transformation, WDAF also facilitates an efficient exploitation of the spatial nonwhiteness mentioned in the previous subsection.

Recently, a related approach using data-based estimation of optimal decoupling transformation matrices was proposed in [137]. Advantages of the data-based approach are that the resulting transformations account for arbitrary array geometries and reverberant environments.

4.30.5 Nonlinear modeling and cancellation of echo

The previous approaches for acoustic echo cancellation assumed that the acoustic echo path can be modeled by a linear system. In practice, however, many loudspeaker systems involve non-negligible nonlinearities, e.g., caused by overloaded amplifiers or low-cost loudspeakers driven at high volume [138,139]. Beyond a certain degree of nonlinear distortion, purely linear approaches are not able to

Author's personal copy

provide sufficient echo attenuation and nonlinear approaches become desirable. Therefore, we recall different approaches that have been proposed for coping with nonlinear acoustic echoes based on simplified models of the acoustic echo path. In Section 4.30.5.1, we prepare the discussion by first outlining physical properties of nonlinear audio hardware components that are typical in hands-free or mobile communication. Then, in Sections 4.30.5.2–4.30.5.4, we present different nonlinear adaptive structures for application in the nonlinear echo cancellation context. Thereby, we distinguish between memoryless nonlinearities such as saturation characteristics of amplifiers on the one hand, and nonlinearities with memory as required for nonlinear loudspeakers on the other hand. In the first case, nonlinear cascaded structures [140] and power filters [141] are considered, whereas in the latter case second-order Volterra filters are of interest [142,143].

4.30.5.1 Nonlinear acoustic echo paths

The structure of an acoustic echo path with possible nonlinearities is illustrated in Figure30.15. As can be seen, it basically consists of the cascade of the amplifier, loudspeaker, and microphone. Additionally, it comprises the acoustic propagation path between the loudspeaker and the microphone.

The propagation path between loudspeaker and microphone can usually be considered as a linear system. It is commonly modeled by a linear FIR filter representing the room impulse response. The microphone signals that are common with hands-free and mobile telephony have only moderate excitation levels. Thus, it is reasonable to assume a linear behavior for the microphone, too, which is in accordance with the observations reported in [140]. We then consider two main sources for nonlinear distortion: The amplifier and the loudspeaker.

Amplifier nonlinearities are especially present in mobile communication devices. There, the dilemma arises to provide high signal levels while having a low battery voltage. Consumers usually prefer an overloading of the amplifier over a reduction of the sound volume. The nonlinear behavior of amplifiers can therefore be described as a memoryless saturation characteristic with a soft clipping of large amplitude values [140].

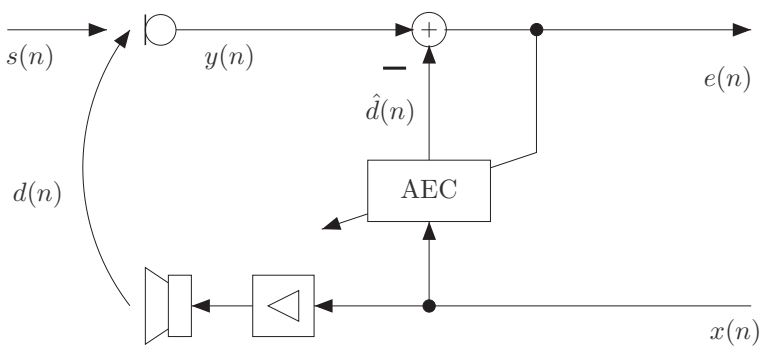


FIGURE 30.15

Hardware setup of the acoustic echo cancellation problem.

Author's personal copy

Many researchers have worked on the characterization of the nonlinearities of electrodynamic loudspeakers (see, e.g., [144,145]). Summarizing their results, one can identify two sources of nonlinear distortion that are relevant in the AEC context: The nonlinearities in the electromagnetic part are mainly caused by the asymmetries of the magnetic flux and its decay outside the air gap of the motor. Thus, the driving force on the voice coil is a nonlinear function of its position. Additionally, in the mechanical part, the nonlinear dependency of the stiffness of the spider and the outer rim on the position of the voice coil has to be taken into account. Without looking at further details, we exploit the main result of [144,145] which imply that the nonlinear behavior of loudspeakers can be modeled by an appropriate Volterra filter. More precisely, we follow [142] and consider the loudspeaker as a black box, whose input/output relation can be approximated sufficiently well by a second-order Volterra filter.

It should also be mentioned that the results presented in [141] indicate that for mobile phones, a saturation-type behavior of the miniaturized loudspeakers can be expected. In this case, soft clipping characteristics as in case of overloaded amplifiers represent a better model.

Other sources for nonlinear distortion in the acoustic echo path can be rattling and vibration effects caused by a strong physical coupling between loudspeaker, microphone, and their enclosure, as, e.g., common in mobile phones. However, this distortion can hardly be modeled or predicted, as it is of chaotic nature [102]. It should rather be considered as uncorrelated noise (analogously to any background noise) and, thus, be processed accordingly. The problem of vibrating system components is not further considered here.

4.30.5.2 Cascaded structure

First, we look at the case where the nonlinear distortion is introduced by an overloaded amplifier. From the discussion in Section 4.30.5.1 it follows, that a simplified model of the nonlinear echo path is given by the cascade of a memoryless saturation characteristic, corresponding to the amplifier, followed by a linear FIR filter. Here, the linear filter corresponds to the remaining propagation path of the echo signal including the loudspeaker, the room, and the microphone. The real nonlinear echo path and its cascaded model is shown on the left hand side and the center of Figure30.16, respectively. The parallel structure depicted on the right hand side will be considered later in Section 4.30.5.3.

The following discussion of adaptive realizations of the cascaded model according to Figure 30.16 center is based on [140]. The input/output relation of the memoryless nonlinearity can be expressed by

$$x_{\mathrm{nl}}(n) = \mathrm{f}(\mathbf{a}, x(n)), \tag{30.90}$$

where $x(n)$ is the input signal and $\mathbf{a}$ denotes a parameter vector that includes all model parameters required to specify the function $f(\cdot)$. Note that since $f(\cdot)$ represents a memoryless nonlinearity, its output $x_{\mathrm{nl}}(n)$ depends only on the current input value $x(n)$. In general, $f(\cdot)$ could be any function that properly models the desired saturation behavior. Possible functions are hard-clipping characteristics [140], or other parametric functions as proposed in [139].

Another general class of memoryless nonlinearities is given by truncated Taylor series expansions and has already been successfully applied to nonlinear AEC in [140]. In case of a Taylor series expansion

Author's personal copy

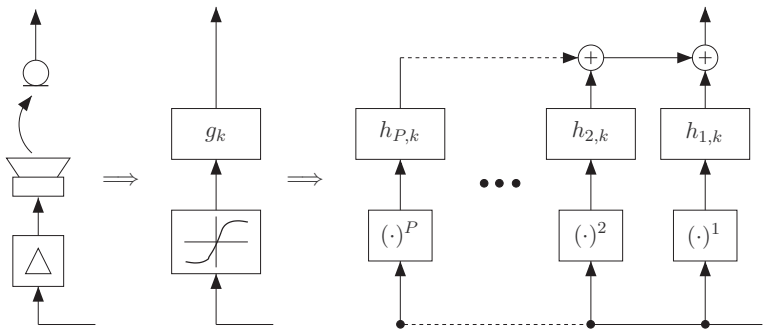


FIGURE 30.16

Illustration of the nonlinear acoustic echo path (left), its cascaded model (center), and a parallelized model with power filters (right).

truncated at order P, Eq. (30.90) becomes

$$x_{\mathrm{nl}}(n) = \sum_{p=1}^{P} a_p x^p(n), \tag{30.91}$$

where a_p represent the coefficients of the Taylor series expansion, i.e., $\mathbf{a} = [a_1, a_2, \ldots, a_P]$ here. To give a practical example, in [140] an order P of seven has been proposed.

The overall output of the cascaded structure is obtained as the linear convolution of $x_{\mathrm{nl}}(n)$ with the linear FIR filter g_k. If the AEC in Figure 30.15 is realized accordingly, the estimate of the echo signal $\hat{d}(n)$ is given by

$$\hat{d}(n) = \sum_{k=0}^{N_g-1} g_k x_{\mathrm{nl}}(n-k). \tag{30.92}$$

Since both, the coefficients of Taylor series expansion a_p, and the coefficients of the linear filter g_k are not known in advance and, moreover, vary in time, they have to be realized adaptively.

The most prominent adaptive algorithm in the AEC context is given by the least mean square (LMS) algorithm [10]. We seek to minimize the mean square of the error signal $e(n)$ at the output of the AEC, where

$$e(n) = y(n) - \hat{d}(n), \tag{30.93}$$

as shown in Figure 30.15. In the following we briefly present the corresponding update equations without further discussions. For more algorithmic details and rigorous derivations, the interested reader is referred to [139,140].

Analogously to linear AEC, the update of the filter coefficients g_k using the LMS algorithm yields

$$g_k(n+1) = g_k(n) + \mu_g(n) e(n) x_{\mathrm{nl}}(n-k), \tag{30.94}$$

Author's personal copy

where $x_{\mathrm{nl}}(n)$ is the output of the memoryless nonlinearity according to (30.91). The corresponding normalized LMS (NLMS) is obtained by normalizing the step-size parameter $\mu_g(n)$ according to

$$\mu_g(n) = \frac{\alpha_g(n)}{\sum_{k=0}^{N_g-1} x_{\mathrm{nl}}^2(n-k)}. \tag{30.95}$$

The normalized step-size parameter $\alpha_g(n)$ is chosen according to $0 < \alpha_g(n) < 2$ to assure stable convergence [10]. To obtain robust adaptation in practice, the step-size parameter has to be controlled to account for distortions such as background noise or double-talk situations [2].

The LMS-type adaptation of the Taylor series expansion is performed according to

$$a_p(n+1) = a_p(n) + \mu_{a_p}(n) e(n) u_p(n), \tag{30.96}$$

where the auxiliary signal $u_p(n)$ is defined as

$$u_p(n) = \sum_{k=0}^{N_g-1} g_k x^p(n-k). \tag{30.97}$$

The normalization of the step-size parameter $\mu_{a_p}(n)$ is given analogously to $\mu_g(n)$ and obtained by replacing $x_{\mathrm{nl}}(n)$ with $u_p(n)$ in (30.95).

In order to increase the convergence speed of the coefficients of the Taylor series expansion, the authors of [140] perform their adaptation via RLS. Since the number of coefficients P is generally small (e.g., $P \leq 7$), the increase in computational complexity compared to the LMS algorithm is rather small, while the convergence speed is significantly improved. A recursive Bayesian algorithm for coupled estimation of a_p and g_k, the *variational Bayesian state-space frequency-domain adaptive filter*, was then proposed as a further update featuring inherent step-size control [146].

It should be noted that miniaturized loudspeakers of mobile phones, when driven into saturation, show similar behavior to overloaded amplifiers [141]. Thus, the above considerations analogously apply in this case.

4.30.5.3 Power filters

The application of cascaded structures that match the model of the nonlinear acoustic echo path, as discussed in the previous section, represents a straightforward and computationally efficient approach to nonlinear AEC. However, it is often challenging to assure convergence to the optimum solution or even assure stable adaption behavior for adaptive *cascaded* structures. In this section, we, therefore, consider so-called power filters as a practical *parallelized* model of the nonlinear echo path in case it includes memoryless saturation characteristics.

The general structure of power filters is illustrated on the right hand side of Figure 30.16. Assuming that the AEC in Figure 30.15 is realized as a Pth-order power filter, its input/output relation reads as follows

$$\hat{d}(n) = \sum_{p=1}^{P} \sum_{k=0}^{N_p-1} h_{p,k} x^p(n-k). \tag{30.98}$$

Author's personal copy

From (30.98) we notice that power filters can be considered as linear multiple input/single output systems, where the input of the pth channel is given by the pth power of $x(n)$. The input of each channel is then filtered by an associated linear filter $h_{p,k}$ with memory length N_p.

As already indicated in Figure 30.16, there is a close relation between the output of the cascaded structure according to (30.92) and the corresponding power filter: Substituting the definition of $x_{\mathrm{nl}}(n)$ according to (30.91) into (30.92) gives

$$\hat{d}(n) = \sum_{p=1}^{P} \sum_{k=0}^{N_g-1} a_p g_k x^p(n-k). \tag{30.99}$$

Comparing (30.98) and (30.99), the power filter model of the corresponding cascaded structure is directly obtained by

$$h_{p,k} = a_p g_k. \tag{30.100}$$

It should be noted, that the number of parameters is increased from $P + N_g$ for the cascaded structure to PN_g for the parallel structure. In practice, however, the increase in number of coefficients is usually much less, as the higher-order channels require less memory compared to the linear channel, i.e., $N_p < N_g$ for $p > 1$. The results reported in [141,143] indicate that power filters of order three already achieve a remarkable increase in echo attenuation compared to linear approaches in case of both, an overloaded amplifier and the nonlinear loudspeaker of a mobile phone.

As already mentioned, power filters can be considered as linear multichannel systems. Thus, a corresponding adaptive realization is straightforward. Here, we only present the LMS algorithm for the update of the coefficients of the power filter $h_{p,k}(n)$, i.e.,

$$h_{p,k}(n+1) = h_{p,k}(n) + \mu_p(n) e(n) x^p(n-k). \tag{30.101}$$

In practice, a control of the step-size parameter $\mu_p(n)$, as well as an appropriate normalization, is important to achieve a reasonable compromise between convergence speed and robustness against distortions such as background noise and double-talk. In nonlinear echo cancellation, the adaptation control additionally has to take into account the influence of nonlinear distortion. A corresponding step-size control and normalization for adaptive power filters has been presented in [143,147].

Again referring to the multi-channel interpretation of power filters, we recall that the input signals of the different channels, i.e., $x(n), x^2(n), \ldots, x^P(n)$ are in general correlated. This implies that the convergence speed of a respective adaptive implementation is rather slow. In order to increase the convergence speed of adaptive power filters, it has been proposed in [141,143] to use corresponding orthogonalized structures instead. The new set of mutually orthogonal input signals for each channel of the power filter is then given by

$$x_{\mathrm{o},p}(n) = x^p(n) + \sum_{i=1}^{p-1} q_{p,i} x^i(n), \tag{30.102}$$

for $1 < p \leq P$, while the linear channel remains unchanged. The orthogonalization coefficients $q_{p,i}$ are chosen such that the cross-correlation between the input signals of different channels becomes zero.

Author's personal copy

The orthogonalization coefficients $q_{p,i}$ can be determined, e.g., by using the Gram-Schmidt orthogonalization method. More details about time-variant orthogonalization for non-stationary input signals such as speech are discussed in [141,143].

In linear adaptive filtering, frequency-domain approaches are known to increase convergence speed while even decreasing computational complexity. Due to their close relation to linear multichannel filtering, an efficient implementation of adaptive power filters in the frequency domain is well possible, as it has been discussed in [143,147]. This is achieved by performing the time-domain update Eq. (30.101) as well as the computation of the output signal $\hat{d}(n)$ (30.98) in the frequency domain. A multichannel recursive Bayesian learning algorithm, the *multichannel state-space frequency-domain adaptive filter* was recently proposed as a contained adaptive algorithm for the power-filter model at hand [148]. It provides inherent stepsize control according to its underlying Kalman filter architecture and has thus proven robustness for noisy and time-varying acoustic environments.

4.30.5.4 Second-order Volterra filters

For the case that the medium-sized loudspeaker of a hands-free telecommunication device represents the main source for nonlinear distortion in the echo path, it has to be modeled by a nonlinearity with memory. As already mentioned in Section 4.30.5.1, second-order Volterra filters represent a suitable model for nonlinear loudspeakers which has already been applied in [142,143].

Assuming that the AEC is realized as a second-order Volterra filter, the AEC output $\hat{d}(n)$ can be expressed by the sum of the output of its linear kernel $\hat{d}_1(n)$, and the output of its quadratic kernel $\hat{d}_2(n)$, i.e.,

$$\hat{d}(n) = \hat{d}_1(n) + \hat{d}_2(n), \tag{30.103}$$

where the input/output relation of the linear and the quadratic kernel, respectively, are given by

$$\hat{d}_1(n) = \sum_{k=0}^{N_1-1} h_k x(n-k), \tag{30.104}$$

$$\hat{d}_2(n) = \sum_{k_1=0}^{N_2-1} \sum_{k_2=k_1}^{N_2-1} h_{k_1,k_2} x(n-k_1) x(n-k_2), \tag{30.105}$$

respectively.

Analogously to the previous sections, we now look at a simplified model for the nonlinear acoustic echo path as shown in Figure 30.17. Assuming that the amplifier of the loudspeaker is sufficiently linear, the echo path can be modeled by the cascade of a second-order Volterra filter $(\widetilde{h}_k, \widetilde{h}_{k_1,k_2})$ representing the nonlinear loudspeaker, followed by a linear filter, corresponding to the room impulse response g_k. This cascaded Volterra structure is illustrated in the center of Figure 30.17. It should be mentioned here, that in [149] the authors propose to realize the nonlinear AEC analogously to such a cascaded Volterra structure. On the one hand, this approach has a rather low computational complexity. However, on the other hand, such adaptive implementations of cascaded systems with memory in general exhibit severe convergence problems, making their application in echo cancellation inappropriate. In the following we thus consider the structure on the right-hand side of Figure 30.17, which consists of a single second-order Volterra filter (h_k, h_{k_1,k_2}) modeling the complete acoustic echo path, i.e., including the room propagation path.

Author's personal copy

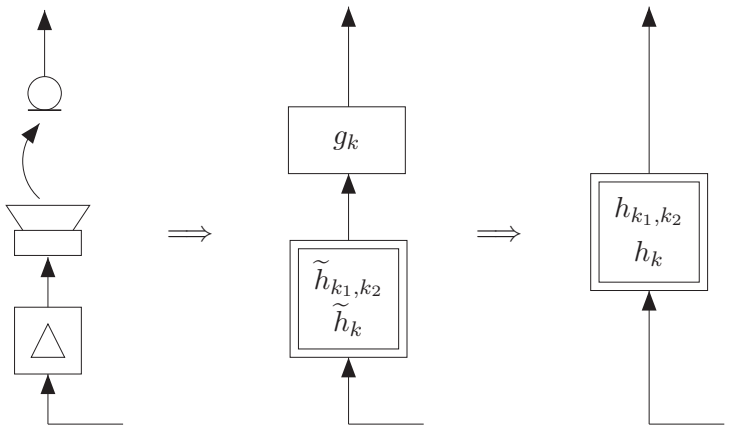


FIGURE 30.17

Illustration of the nonlinear acoustic echo path (left), its cascaded Volterra filter model (center), and the corresponding overall model of a single Volterra filter (right).

As has been shown in [143,150], the corresponding quadratic Volterra kernel has specific properties, namely a large part of the coefficients of the quadratic kernel are known to be zero in advance.

In order to exploit the a priori knowledge about the kernel coefficients for efficient implementations, it is useful to employ an alternative representation of the quadratic kernel. Following [150], we rewrite (30.105) using the so-called diagonal coordinate representation (DCR):

$$\hat{d}_2(n) = \sum_{r=0}^{R-1} \sum_{k=0}^{N_2-r-1} h_{k,r+k} x(n-k) x(n-r-k). \tag{30.106}$$

Comparing (30.105) and (30.106), we notice that the above computation of $\hat{d}_2(n)$ can be interpreted as the summation over R diagonals within the Cartesian coordinate system constructed by the summation indices (k_1, k_2). Thereby, the main diagonal corresponds to $r = 0$. Obviously, in case of $R = N_2$, (30.105) and (30.106) are equivalent.

Referring to the cascaded structure shown in the center of Figure 30.17, we now assume that the leading Volterra filter has a memory length of L for both, linear and quadratic kernel. Furthermore, we assume that the linear filter g_k has length N_g. As shown in [150], the resulting overall Volterra filter has memory lengths $N_1 = N_2 = L + N_g - 1$. However, the width R of the corresponding DCR of the quadratic kernel remains unchanged, i.e., $R = L$. Considering that L represents the memory effects of the loudspeaker and N_g corresponds to the reverberation time of the room, it becomes obvious that in typical applications $R \ll N_2$. As can be seen from (30.106), this special property can easily be taken into account when using the DCR of Volterra filters.

Author's personal copy

Another interesting property of Volterra filters can be found when introducing the virtual input signal $x_r(n) = x(n)x(n-r)$ of the rth diagonal into (30.106)

$$\hat{d}_2(n) = \sum_{r=0}^{R-1} \sum_{k=0}^{N_2-r-1} h_{k,r+k} x_r(n-k). \tag{30.107}$$

As can be seen, the inner summation represents a linear convolution between the kernel coefficients on the rth diagonal with the input signal $x_r(n)$. Thus, quadratic Volterra kernels can be considered as a special type of linear multichannel systems, where each diagonal of the DCR corresponds to one channel with input $x_r(n)$. Regarding this, algorithms known from linear adaptive filtering can easily be extended to adaptive second-order Volterra filters. To give an example, the update equation for the coefficients of the quadratic kernel using the LMS algorithm is given by

$$h_{k,r+k}(n+1) = h_{k,r+k}(n) + \mu_{k,r+k}(n) e(n) x_r(n-k). \tag{30.108}$$

A detailed discussion of suitable methods for the normalization and control of the step-size parameter $\mu_{k,r+k}(n)$ can be found in [143].

Due to the close relation between Volterra filters and linear multichannel systems, the derivation of corresponding efficient frequency-domain realizations is straightforward. For instance in [143,150], it has been proposed to perform the linear filtering required for each diagonal in (30.106) by using fast block convolution techniques in the frequency domain. Additionally, the update of the kernel coefficients $h_{k,r+k}(n)$ can also be performed in the frequency domain. It turns out, that the benefits of frequency-domain approaches as known from linear adaptive filtering also transfer to adaptive Volterra filters.

Apart from the described potential of nonlinear modeling and identification using Volterra filters, the huge computational complexity and slow convergence related to the large number of parameters have been clearly recognized as limitations regarding the usability. As a result, significant research has been devoted recently to the design of fast and robust algorithms using iterated coefficient update [151] and to complexity reduction via dynamical adjustment of the kernel memory [152].

4.30.6 Application to realistic and real systems

In this section, we describe various acoustic environments with different configuration regarding the hands-free communication application. In particular, we consider the car, the desktop PC, the living room, and the mobile phone environment. Essentially, these environments exhibit individual degrees of environmental noise, length and time-variability of the acoustic echo path, and nonlinearities such as sampling asynchrony or loudspeaker saturation. As a consequence, different signal processing approaches have been used by researchers to tackle the acoustic echo control problem in the different environments. In the following, the results that have been achieved are outlined along with main properties of the respective environment.

4.30.6.1 Car environment

Due to the relatively small size acoustic environment of the car interior, we have a relatively short echo path impulse response of only 30–100 ms duration in most of the cases. However, the natural presence

Author's personal copy

and interaction of the user in the environment will cause the echo path impulse response to exhibit relatively strong variability, which practically means a lack of identifiability if at the same time the natural presence of the car noise is considered. Regarding the quality of electro-acoustic transducers, at least in the high-end product range, we may assume only a minor degree of nonlinearity of the system. As a result, the linear state-space echo path model and the respective model-based optimum filtering approaches as described in Section 4.30.2 were found to best address this environment. In order to reach out for the limits, here, we evaluate the advanced implementation of the frequency-domain adaptive Kalman filter as proposed in [27].

In order to allow for reproducibility of the presented results, while maintaining strong relationship with the real-world situation, we make use of a time-varying echo path that is generated directly by the Markov model in (30.34). The variability is chosen such that the echo attenuation of a perfectly adjusted echo canceler would drop to about 0 dB within 2–3 s after the adaptation of the filter is halted. The echo path vector $\mathbf{h}(n)$ contains 512 coefficients, which corresponds to 64 ms echo path duration at 8 kHz sampling frequency. To setup the test signals for the adaptive algorithm, we use real speech input on both the far-end and near-end side of the communication. The employed speech material consists of 8 phonetically balanced sentences (male and female) of about 5 s duration each [153]. We then consider a wide range of signal-to-echo ratios $\mathrm{SER}_y = \sigma_s^2/\sigma_d^2$ at the hands-free microphone. The $\mathrm{SER}_y = 0$ dB simulates a hard double talk situation, $\mathrm{SER}_y = -40$ dB corresponds to remote single talk, and $\mathrm{SER}_y = 40$ dB finally represents near-end single talk. The background noise level at the hands-free microphone is adjusted such that the signal-to-noise ratio of the near-end speech is 10 dB, while the received signal from the far-end speaker is almost clean speech with a signal-to-noise ratio of 40 dB—a situation that often exists in car hands-free communication.

The *echo attenuation* after echo canceler and postfilter, cf. (30.28) and (30.29), can be evaluated in terms of the echo return loss enhancements $\mathrm{ERLE}_{w_1} = \sigma_d^2/\sigma_b^2$ and $\mathrm{ERLE}_{w_{12}} = \sigma_d^2/\sigma_{b'}^2$. Here, $b = d - \hat{d}$ refers to the residual echo after echo cancellation and $b' = w_2 * b$ represents the total echo attenuation after both filters, e.g.,[2,58]. The resulting *speech quality* is evaluated by means of the resulting signal-to-echo ratio $\mathrm{SER}_e = \sigma_s^2/\sigma_{s-e}^2$ after the echo canceler and $\mathrm{SER}_{\hat{s}} = \sigma_s^2/\sigma_{s-\hat{s}}^2$ at the system output, i.e., after the postfilter. The ERLE and SER measures described here are suitable to characterize the overall performance of echo canceler and postfilter, including the adaptive algorithm with its tracking performance and robustness against observation noise.

When echo canceler and postfilter and the adaptive algorithm (i.e., the frequency-domain adaptive Kalman filter) are implemented with a block frame-shift (i.e., algorithmic delay) of 8 ms and a DFT size of 512 (corresponding to 64 ms echo path impulse response length), and when the time-constant of the Kalman filter is matched to the dynamical echo path model, we obtain the results in Figure 30.18. The ERLE_{w_1} by the echo canceler ranges from 0 to 20 dB, depending on the input SER_y. The saturation of ERLE_{w_1} at low SER_y is due to the time-varying echo path and the fact that the echo canceler for time n is determined by the "incomplete" data available up to time $n-1$. For high SER_y, ERLE_{w_1} asymptotically reaches zero, since extremely noisy observations do not allow the identification of the time-varying echo path at all. The total echo attenuation $\mathrm{ERLE}_{w_{12}}$ by echo canceler and postfilter ranges from 0 to 50 dB. This performance matches the industrial requirements for acoustic echo controllers: More than 40 dB ERLE is indeed recommended during remote single talk [78,84]; in noisy double talk situations our experience is that 15–20 dB ERLE is sufficient to achieve

Author's personal copy

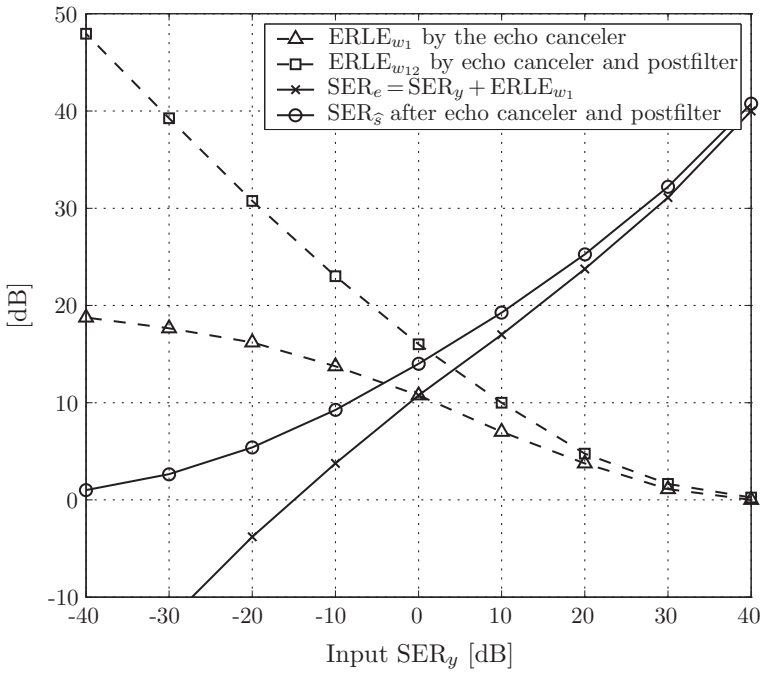


FIGURE 30.18

ERLE and output SER for different input SER.

the required end-user quality; and during near-end single talk an echo attenuation is of course not required.

The speech quality improvement by the echo canceler can then be expressed analytically: $\text{SER}_e = \text{SER}_y + \text{ERLE}_{w_1}$, thus $\text{SER}_e > \text{SER}_y$. The situation is not so straightforward in case of the postfilter, but from Figure 30.18 we observe another consistent improvement in the output SER, i.e., $\text{SER}_{\hat{s}} > \text{SER}_e$. For very low input SER_y, a surprisingly high output $\text{SER}_{\hat{s}} \approx 0$ dB is attained, simply because the entire microphone signal is strongly attenuated. For high input SER_y, the output $\text{SER}_{\hat{s}}$ approaches the input SER_y since the microphone signal remains nearly unprocessed. For $\text{SER}_y = 0$ dB, i.e., during double talk, we have $\text{SER}_{\hat{s}} \approx 14$ dB. However, together with the effect of perceptual masking, the perceptual signal quality ("the perceived SER") is much better than $\text{SER}_{\hat{s}} \approx 14$ dB.

4.30.6.2 Desktop conferencing

Based on computers connected to the Internet, a widespread hands-free telecommunication application is desktop conferencing. To set up a desktop conference call, one only needs a computer connected to the Internet, a loudspeaker, a microphone (and potentially a camera for display) as shown in Figure30.19. In order to allow hands-free calls, the computer requires a software carrying out echo control. Some

Author's personal copy

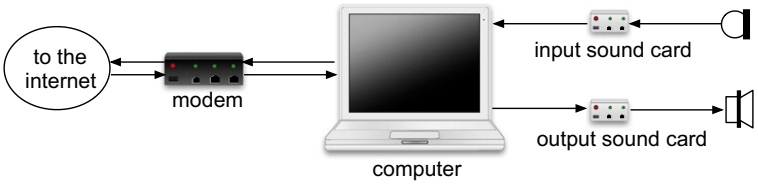


FIGURE 30.19

A general desktop conferencing environment.

operating systems have the required software application already pre-installed. Otherwise, users can easily download from the Internet any available conferencing software.

The variety of computers and hardwares (loudspeakers, microphones, sound cards, etc.) makes the design of the echo control software a challenging task. Indeed, the echo control software has to work despite the various possible computer architectures and operating systems, speaker sizes and efficiencies, microphones types and sensitivities, etc., and even further, the various and uncontrollable user environments. In order to cope with these numerous unknowns, a robust acoustic echo suppression (AES) system as described in Section 4.30.3 can be used. It offers a viable solution, since it does not require an exact identification of the impulse response h_n, but models parametrically the echo path with a delay τ and a single real-valued gain $\widehat{G}(m, \nu)$ at each frequency bin of short-time spectra as shown by Eq. (30.59). Therefore, AES yields robust insurance against movements of the microphone or other changes in the acoustic environment.

Another important feature of the AES is that the phase information of the signal spectra is discarded in the algorithm, making the AES performance independent of any phase changes or distortions introduced by the components in the acoustic echo path. In the specific desktop conferencing environment, the two most common distortions are:

- *Sampling rate mismatch:* Leading to time drift, which typically arises when the loudspeaker signal and the microphone signal are captured using different soundcards or A/D converters.
- Random loss of audio samples or frames of samples due to transmission over the IP network or due to drop outs during playback.

As already illustrated in Figure 30.6, the described AES implementation is robust against such common issues. The estimate of the echo estimation filter (EEF) function $\widehat{G}(m, \nu)$ is computed directly from power spectra with temporal fluctuations instead from complex spectra, which not only makes the estimate insensitive towards phase distortions, but also makes it independent to the background noise on the near-end side as seen from Eq. (30.71).

Eventually, the concept of AES, based on a spectral subtraction of the echo estimate from the microphone spectrum, enables to compute "aggressively" the final echo suppression filter (ESF) $F(m, \nu)$ such that no residual echo remains. This can be achieved by choosing a long reverberation time constant α_{RT} in Eq. (30.72) to match the room size and suppress the late echoes, or by intentionally over-estimating the echo power spectrum with a large β parameter in Eq. (30.74). Because the ESF can be adjusted to

Author's personal copy

perform more aggressive echo suppression, independent from the estimation of the EEF, a tuning point can be found where the AES also provides a certain insensitivity against non-linear behavior of the echo path.

In summary, for the specific desktop environment, AES provides a flexible approach to perform echo control. It is a practical solution to cope with the variety of possible hardwares and related system uncertainties.

4.30.6.3 Living room

In order to demonstrate the potential of multichannel acoustic echo cancellation (MCAEC), as described in Section 4.30.4, in applications such as home theater, virtual reality, or advanced teleconferencing, we chose the acoustic environment of a typical living room and consider a surround sound scenario with $P = 5$ reproduction channels. The sampling rate of the loudspeaker signals and the preprocessing stage is 44.1 kHz, while the microphone signal and the echo cancellation is downsampled by a factor of 4 as typical for speech recognition applications. The length of the echo cancellation filters were set to $N = 1024$, covering the reverberation time in the receiving room. Our evaluation mostly relies on the MCFDAF algorithm [26], which aims to exploit all cross-correlations between the reproduction channels. Besides the iterative processing in time, our implementation performs 10 offline iterations within each block of samples according to [55,56]. In all simulations, the echo-to-background noise ratio in the receiving room was set to 30 dB and the regularization of the MCFDAF algorithm was adjusted so that stability is provided for all preprocessing methods under investigation.

At first, we discuss the convergence of the adaptive filter coefficients to the true echo path coefficients in terms of the multichannel coefficient error norm $\sum_{i=1}^{P} \|\mathbf{h}_i - \hat{\mathbf{h}}_i(n)\|^2 / \sum_{i=1}^{P} \|\mathbf{h}_i\|^2$ over time with different preprocessing methods. We chose a somewhat critical test scenario of reproducing a narrowband high-quality male speech signal with alternating spatial positions in the transmission room (see Figure 30.11). In order to reflect the surround sound scenario and the inherent level imbalance problem in MCAEC appropriately, it is important to choose a realistic recording scenario, ours being inspired by the so-called Decca Tree and surround microphones [154]. Figure30.20 then shows the corresponding coefficient convergence for baseline approaches without any preprocessing (curve label "without preproc.") and with conventional nonlinear preprocessing after [115] (nonlinearity parameter $\alpha = 0.5$, label "NL"), as well as for the perceptually tuned frequency selective phase modulation method [120] (label "Pmod_fs") and the addition of uncorrelated audio coding noise after [128] (labeled "mp3_48"). As it can be seen from the data, convergence without any preprocessing is extremely slow, while preprocessing results in a significant convergence boost. The parameters of all preprocessing methods considered here were chosen such that they yield similar convergence characteristics in order to provide a common basis for subjective listening tests, such as the MUSHRA in Section 4.30.4.3.

Secondly, by choosing the phase modulation method as a fixed preprocessor, we illustrate the effect of taking into account the cross-correlations between the loudspeaker channels in the AEC coefficient update (see also Figure 30.12 in Section 4.30.4.2). We again apply the MCFDAF algorithm for $P = 5$ loudspeaker channels with the same parameters and the same data as above and then draw the comparison with a standard FDAF algorithm in each and every channel, specifically, the unconstrained fast least mean-square (UFLMS) algorithm. The results in Figure 30.21 clearly confirm the significant convergence improvement regarding ERLE and coefficient error norm by taking into account the cross-correlations into the adaptation process.

Author's personal copy

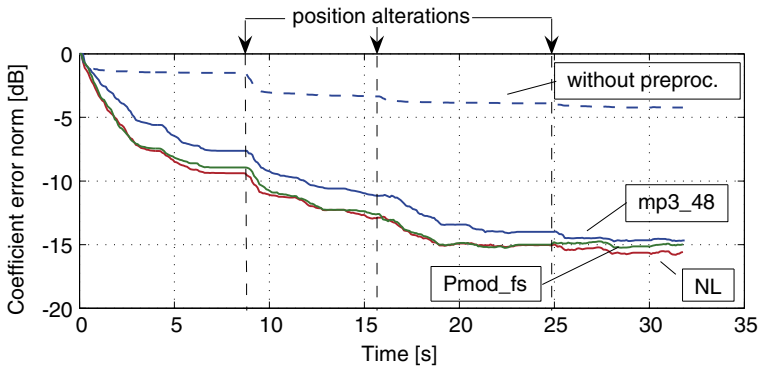


FIGURE 30.20

Comparison of MCAEC processing methods, $P = 5$.

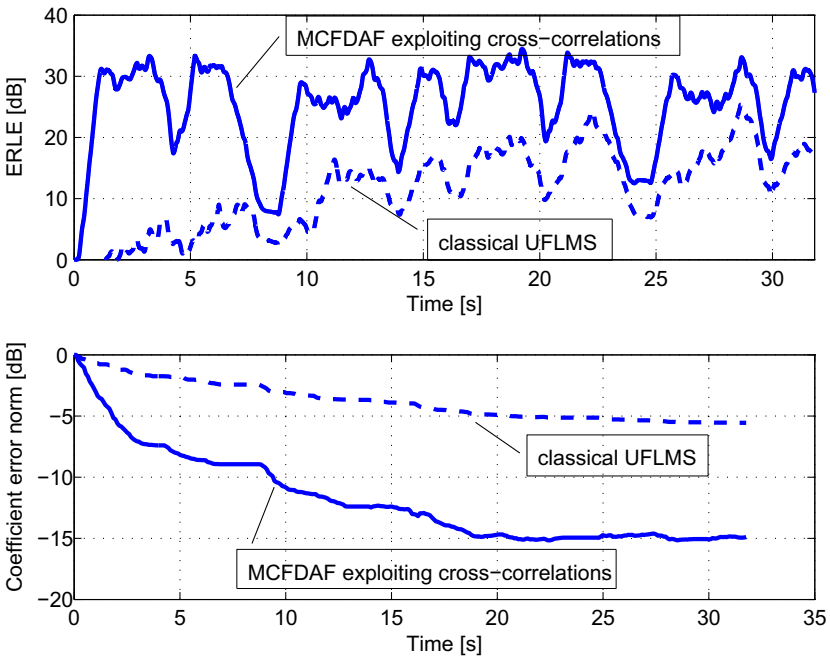


FIGURE 30.21

Effect of taking cross-correlations into account, $P = 5$ channels. (a) ERLE convergence, (b) coefficient error norm.

Author's personal copy

4.30.6.4 Mobile phones

In Section 4.30.5, we discussed nonlinear echo path models and adaptive filter structures which require only little *a priori* knowledge about the audio hardware employed in telecommunication devices. If moderately-sized loudspeakers represent the only source of nonlinear distortion, then second-order Volterra filters are generally recommended to model their frequency-dependent nonlinear behavior. In case of memoryless nonlinearities included in the echo path, as common with nonlinear amplifiers or miniaturized loudspeakers of mobile phones, the nonlinear cascaded structures including truncated Taylor series expansions or, alternatively, power filters are better suited. In this section, we explore the suitability of all these approximations when modeling the real acoustic echo path comprising the miniature electro-dynamic loudspeaker of a mobile phone.

For recordings of audio signals, the loudspeaker has been mounted in the handset, while the microphone has been separated from it to avoid undesired vibration effects due to physical coupling of the loudspeaker and the microphone. During the measurements it has been assured that there is no nonlinear distortion introduced by overloading of the amplifier, i.e., the nonlinearity in the acoustic echo path is mainly caused by the miniature loudspeaker. In order to focus on the nonlinear behavior of the setup, the recordings have taken place in a room with low reverberation. The input signal has been wide-sense stationary correlated Gaussian noise, which has been generated by passing a white Gaussian noise signal through a second-order recursive filter.

First, we evaluate the suitability of the different nonlinear structures for modeling the nonlinear behavior of the loudspeaker. The behavior is examined by applying three different input levels to five different adaptive structures. The considered structures are: A linear filter, a third- and fifth-order orthogonalized power filter, and a second- and third-order Volterra filter in DCR, respectively. All approaches have been implemented in the DFT domain to improve the convergence properties for correlated input. Since the input signal used for the measurements is known in advance, a fixed orthogonalization of the channel inputs can be used for the power filters. The memory length of the linear filter and the linear channel of the nonlinear approaches has been $N_1 = 300$ taps, which sufficiently models the linear component of the echo path. The filters associated with the nonlinear channels of both, third-order and fifth-order power filter have been implemented with a length of $N_p = 100$ taps. Accordingly, the memory lengths of the nonlinear kernels of second- and third-order Volterra filters have been set to $N_2 = N_3 = 100$. Accounting for the cascaded structure of the acoustic echo path, the widths of the quadratic and cubical kernels have been reduced to $R_2 = R_3 = 10$.

The evaluation of the different approaches is based on the maximum ERLE that is achieved after convergence of the echo canceler. The resulting final ERLE values obtained for different input variances are summarized in Table 30.1. The first column corresponds to the case where there is only a low level of nonlinear distortion in the echo path. This is reflected by the fact that the linear adaptive filter shows approximately the same performance as the nonlinear counterparts. Thereby, we notice that the achievable ERLE values are generally not so large. This can be explained by the fact that the relatively low sound level of the miniature loudspeaker allows not more than 30 dB SNR at the recording microphone. In the first row of the table, the ERLE obtained for the linear adaptive filter is reduced by approximately 5–6 dB when the input signal level is increased to $5.4\sigma_{m,x}^2$ and $9\sigma_{m,x}^2$, respectively. This confirms the nonlinear behavior of the loudspeaker for high excitation levels.

Author's personal copy

Table 30.1 Achievable ERLE of Several Adaptive Structures Obtained for Different Input Levels. The Nonlinear Distortion is Introduced by the Loudspeaker of a Mobile Phone

Variance of the input	$\sigma^2_{m,x}$ (dB)	$5.4\sigma^2_{m,x}$ (dB)	$9\sigma^2_{m,x}$ (dB)
Linear filter	27.2	22.3	21.1
Third-order power filter	28.4	25.4	24.4
Fifth-order power filter	28.3	25.4	24.5
Second-order Volterra filter	26.9	22.2	22.1
Third-order Volterra filter	25.9	25.6	25.4

The second-order Volterra filter does not yield noticeable improvements compared to the linear filter. This indicates that the memoryless model for the miniaturized loudspeakers of the mobile phone is sufficient. By extending the second-order Volterra filter to a cubical kernel, the echo attenuation of the linear filter is surpassed by approximately 3–4 dB for the input variances $5.4\sigma^2_{\mathrm{m},x}$ and $9\sigma^2_{\mathrm{m},x}$, respectively. This shows that the nonlinearity of the loudspeaker is at least of third order. Note that the considered third-order kernel requires 5170 coefficients. Additional simulations have shown that the memory length of the nonlinear kernels can be reduced to $N_2 = N_3 = 64$ without changing the maximum achievable ERLE. However, this reduction of the region of support of the third-order Volterra kernel still requires 4070 coefficients. This large number of coefficients and the related difficulty of accurate adaptive identification also explains the performance loss of the third-order Volterra filter that is observed for the lowest input level $\sigma^2_{\mathrm{m},x}$. Regarding that the ERLE gain is at most 4.3 dB for the highest input variance, there is no reasonable relation between performance improvement and increase in computational complexity when applying third-order Volterra filters instead of linear filters.

When the region of support of the third-order Volterra filter only includes the main diagonals of each kernel, it is simplified to a third-order power filter. As can be seen from Table 30.1, this enormous reduction of the region of support barely affects the achievable echo attenuation. This result again supports the assumption that the miniature loudspeaker can be considered as a memoryless nonlinearity. One might expect that increasing the order of the power filter, and thus its nonlinear modeling ability, should then lead to yet more echo attenuation. Unfortunately, an extension of the power filter to fifth order does not yield further improvements over the third-order case in our practical experiments. Our understanding is that the misadjustment of the linear and the cubical channels inhibit the convergence of channels with yet higher orders—in conjunction with the fact that higher order channels of our EOS (equivalent orthogonal structure) are hardly excited.

From the results presented in Table 30.1 we conclude that the modeling capabilities of the considered polynomial filters are not completely satisfying. From a practical point of view, the best compromise with respect to achievable echo attenuation and computational complexity is provided by the orthogonalized third-order power filter. This configuration is therefore used in the following experiment, where we look at the performance of the adaptive EOS of a third-order power filter with real speech input.

Except for the speech input, the experimental setup now is the same as before.The variance of the speech signal has been adjusted such that its amplitude values lie in the same range as the amplitudes of typical sample functions of the correlated noise signal with variance $9\sigma^2_{\mathrm{m},x}$ as used above. A white

Author's personal copy

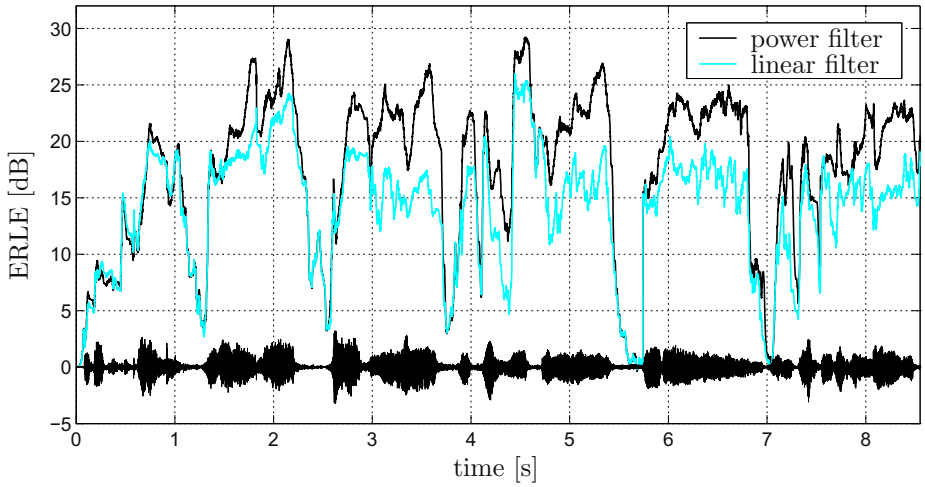


FIGURE 30.22

ERLE obtained for the adaptive EOS of a third-order power filter and a corresponding linear approach together with the speech input.

Gaussian noise signal has been added to the recording of the microphone signal in order to simulate a background noise level corresponding to an SNR of 30 dB with respect to the acoustic echo. Since an algorithmic delay is not desirable in mobile phones, we now consider the time-domain implementation of the EOS, where the memory length $N_1 = 256$ for the linear channel and $N_2 = N_3 = 100$ for both, the quadratic and cubical channel has been chosen. The orthogonalization of the channel inputs has been performed signal-adaptively, where the required moments are estimated recursively with a time-constant of about 10 ms.

In Figure 30.22, the echo cancellation performance of the adaptive EOS of the third-order power filter is compared to a linear approach which corresponds to the linear channel of the power filter. As can be noticed, the performance of the linear adaptive filter is remarkably limited due to the nonlinear distortion introduced by the loudspeaker. The third-order power filter succeeds in improving the level of echo attenuation during almost the whole simulation period. Especially for speech segments that exhibit high excitation levels, the local increase of the ERLE even exceeds the expectations which would have been predicted from Table 30.1. While this ERLE gain is observed, note that due to the short filters in the nonlinear channels, the computational complexity of the considered orthogonalized power filter is only two times higher than that of the linear filter.

4.30.7 Links to codes and recommendations

Echo control solutions have been developed over years and used in many telecommunication applications. Therefore the portfolio of existing solutions is large and includes various implementations as the ones described in Sections 4.30.1–4.30.5 , or all possible combinations to match the application

Author's personal copy

specifications and requirements. Most of the solutions are thus specific and proprietary meaning that available free implementations or data sets are rare and not designed to address diverse applications. Prominent examples of free implementations are:

- An acoustic echo canceler with postfiltering is part of the Speex speech codec (http://www.speex.org).
- The OSLEC line echo canceler (http://www.rowetel.com/ucasterisk/oslec.html).

While implementations of echo control solutions are often proprietary, common knowledge is listed in standards and recommendations referenced by applications and fields of use. The International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T at `http://www.itu.int/ITU-T`) defines for a wide range of possible telecommunication applications a number of recommendations and standards for related echo control solution implementations. For instance:

- `ITU-T G.131:` Talker Echo and its Control [1].
- `ITU-T G.164:` Echo suppressors [75].
- `ITU-T G.165:` Echo cancelers [76].
- `ITU-T P.832:` Subjective performance evaluation of hands-free terminals [81].

The international recommendations produced by the ITU-T can become mandatory once they are adopted as part of a national law to regulate telecommunication applications. Many others standardization organizations have been created to define specifications for echo control solutions in specific fields of use. For examples the 3rd Generation Partnership Project (3GPP at http://www.3gpp.org/) or the German Automobile Industry (VDA at http://www.vda.de/en/index.html), [84], are two organizations writing standards for mobile networks and car applications, respectively.

4.30.8 Conclusions, open issues, future trends

Acoustic echo control for hands-free communication has been actively researched in the area of signal processing since the 1970s. Here, we first reviewed the most popular solutions based on adaptive algorithms according to deterministic least-squares design, with realizations in time- or frequency-domain, and combined with various possible control strategies.

Based on this brief status, our chapter then mainly reported the comprehensive extensions of the state-of-the-art according to research work beyond 2000. This includes statistical methods for adaptive algorithm design, e.g., the unification of adaptive filtering and adaptation control based on statistical echo path modeling and Bayesian estimation, the echo suppression technique based on power spectral echo path modeling, and the TRINICON framework to incorporate statistical signal properties. Furthermore, we highlighted the particular issues of multichannel and nonlinear adaptive systems and the respective developments.

Those new directions in adaptive systems research were triggered by the common need for fast and robust solutions in real-world applications in which we often face a lot of uncertainty regarding the electroacoustic environment and the specific usage of hands-free systems. On these common grounds, however, the new technologies have been conceived and pursued somewhat independently by different researchers in different applications and in different organizations. In this chapter, we have sought a presentation with unified notation, but there remains a lot of work towards a unified and fully

Author's personal copy

comprehensive theory. Nonetheless, our chapter has proven the usability of the presented algorithms in practical single-channel applications, such as desktop and car environments, while the perspective for the realization of systems with multiple reproduction channels and nonlinear characteristics has been demonstrated in the research environment.

Acoustic echo control continuous as a research topic to serve as the enabling technology in modern configurations of hands-free communication with full duplex ability. Future trends include a shift of the acoustic echo control unit from mobile devices into the core of a cellular network in order to save processing power on the mobile device [155,156], the development of multichannel echo cancellation frameworks for systems with massive multichannel reproduction of spatial audio [157], and the evolution of nonlinear adaptive signal processing beyond the established polynomial modeling [158]. In all the cases mentioned here, the derivation of fast and robust adaptive algorithms for the respective structure and system model at hand represents an interesting topic for future research activities.

Glossary

Acoustic echo control	the term generalizes the acoustic echo cancellation to further include echo suppression and postfiltering
Double talk	a situation in which the talkers at both ends of a communication system (or reproduction unit and user of a speech dialog system) are active simultaneously; echo and target input signal are thus superimposed and recorded together at the microphone
Duplex ability	it describes to which degree the simultaneous transmission in reproduction and acquisition direction is preserved by a hands-free system, despite possible attenuations by signal processing
Echo path	this terms describes the undesirable electroacoustic coupling between loudspeaker input and microphone output of a hands-free communication system; most of the times, we describe the echo path in terms of an acoustic impulse response or frequency response comprising loudspeaker unit, acoustic coupling, and microphone
Echo cancellation	refers to a family of techniques, where the echo path is mimicked by an (adaptive) digital filter in order to regenerate and ideally subtract the echo from the observed microphone signal
Echo suppression	in contrast to echo cancellation, this refers to a family of techniques which discard (i.e., not mimic) the phase of the echo path; the echo suppression is then performed in the form of statistical echo reduction based on the echo power transfer function
Multichannel echo cancellation	most of the times, this term refers to the problem of echo cancellation for multiple reproduction channels, e.g., stereophonic echo cancellation
Nonlinear echo cancellation	most of the times, this term refers to the problem of echo cancellation in the presence of a nonlinear power amplifier or nonlinear loudspeaker within the echo path

Author's personal copy

Postfiltering	it describes echo suppression techniques when they are employed in conjunction with echo cancellation
Power filter	a multi-input/single-output, adaptive echo cancellation filter structure with higher-order polynomial representations of the reproduction signal at the multiple inputs; represents a quasi-linear expansion of echo paths with memoryless nonlinearity

Relevant Theory: Signal Processing Theory

See Volume 1, Chapter 3 Discrete-Time Signals and Systems
See Volume 1, Chapter 4 Random Signals and Stochastic Processes
See Volume 1, Chapter 6 Digital Filter Structures and Their Implementation
See Volume 1, Chapter 7 Multirate Signal Processing for Software Radio Architectures
See Volume 1, Chapter 9 Discrete Multi-Scale Transforms in Signal Processing
See Volume 1, Chapter 12 Adaptive Filters

References

[1] ITU-T Rec. G.131, Talker echo and its control, November 2003.
[2] E. Hänsler, G. Schmidt, Acoustic Echo and Noise Control: A Practical Approach, Wiley, 2004.
[3] M. Sondhi, An adaptive echo canceller, Bell Syst. Tech. J. XLVI (3) (1967) 497–511.
[4] M. Sondhi, W. Kellermann, Echo cancellation for speech signals, in: S. Furui, M. Sondhi (Eds.), Advances in Speech Signal Processing, Marcel Dekker, New York/Basel/Hong Kong, 1992, pp. 327–356.
[5] A. Liavas, P. Regalia, Acoustic echo cancellation: do IIR models offer better modeling capabilities than their FIR counterparts, IEEE Trans. Signal Process. 46 (9) (1998) 2499–2504.
[6] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, J. Tilp, Acoustic echo control, an application of very-high-order adaptive filters, IEEE Signal Process. Mag. (1999) 42–69.
[7] S.L. Gay, J. Benesty (Eds.), Acoustic Signal Processing for Telecommunications, Kluwer Academic Publishers, 2000.
[8] E. Hänsler, The hands-free telephone problem – an annotated bibliography, Signal Process. 27(3) (1992) 259–271.
[9] E. Hänsler, The hands-free telephone problem: an annotated bibliography update, Ann. Télécommun. 49 (7–8) (1994) 360–367.
[10] S. Haykin, Adaptive Filter Theory, fourth ed., Prentice-Hall, Upper Saddle River, NJ, 2002.
[11] C. Antweiler, Orthogonalisierende Algorithmen für die digitale Kompensation akustischer Echos, PhD thesis, RWTH Aachen, in: Peter Vary (Ed.), Aachener Beiträge zu digitalen Nachrichtensystemen, Band 1, Verlag der Augustinus Buchhandlung, Aachen, February 1995.
[12] S. Yamamoto, S. Kitayama, J. Tamura, H. Ishigami, An adaptive echo canceller with linear predictors, Trans. IECE Jpn. E62 (12) (1979) 851–857.
[13] K. Ozeki, T. Umeda, An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties, Electron. Commun. Jpn. 67-A (1984) 19–27.
[14] D. Morgan, S. Kratzer, On a class of computationally efficient rapidly converging, generalized NLMS algorithms, IEEE Signal Process. Lett. 3 (1996) 245–247.

Author's personal copy

[15] S. Gay, The fast affine projection algorithm, in: S. Gay, J. Benesty (Eds.), Acoustic Signal Processing for Telecommunications, Kluwer Academic Publishers, 2000, pp. 23–45.
[16] D. Slock, T. Kailath, Numerically stable fast transversal filters for recursive least-squares adaptive filtering, IEEE Trans. Signal Process. 39 (1991) 92–114.
[17] J. Cioffi, T. Kailath, Fast, recursive-least-squares transversal filters for adaptive filtering, IEEE Trans. Acoust. Speech Signal Process. 34 (1984) 304–337.
[18] D. Falconer, L. Ljung, Application of fast Kalman estimation to adaptive equalization, IEEE Trans. Commun. 26 (1978) 1439–1446.
[19] E. Ferrara, Frequency-domain adaptive filtering, in: C. Cowan, P. Grant (Eds.), Adaptive Filters, Prentice Hall, Englewood Cliffs, NJ, 1985, pp. 145–179.
[20] W. Kellermann, Analysis and design of multirate systems for cancellation of acoustical echoes, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New York, April 1988, pp. 2570–2573.
[21] J. Shynk, Frequency-domain and multirate adaptive filtering, IEEE Signal Process. Mag. 9 (1) (1992) 14–37.
[22] E. Moulines, O. Amrane, Y. Grenier, The generalized multidelay adaptive filter: structure and convergence analysis, IEEE Trans. Signal Process. 43 (1) (1995) 14–28.
[23] P. Sommen, Adaptive Filtering Methods: On Methods to Use a Priori Information in Order to Reduce Complexity While Maintaining Convergence Properties, PhD thesis, Technical University of Eindhoven, 1992, ISBN 90-9005143-0.
[24] J.-S. Soo, K. Pang, Multidelay block frequency domain adaptive filter, IEEE Trans. Acoust. Speech Signal Process. 38 (1990) 373–376.
[25] G. Enzner, P. Vary, A soft-partitioned frequency-domain adaptive filter for acoustic echo cancellation, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong (China), May 2003, pp. 393–396.
[26] H. Buchner, J. Benesty, W. Kellermann, Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication, Signal Process. 85 (3) (2005) 549–570.
[27] G. Enzner, P. Vary, Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones, Signal Process. 86 (6) (2006) 1140–1156.
[28] P. Vary, R. Martin, Digital Speech Transmission – Enhancement, Coding, and Error Concealment, John Wiley & Sons, Ltd., Chichester, England, 2006.
[29] R. Merched, P.S. Diniz, M.R. Petraglia, A new delayless subband adaptive filter structure, IEEE Trans. Signal Process. 47 (6) (1999) 1580–1591.
[30] D.R. Morgan, J.C. Thi, A delayless subband adaptive filter architecture, IEEE Trans. Signal Process. 43 (8) (1995) 1819–1830.
[31] D. Duttweiler, Proportionate normalized least-mean-squares adaptation in echo cancelers, IEEE Trans. Speech Audio Process. 8 (2000) 508–518.
[32] J. Benesty, S. Gay, An improved PNLMS algorithm, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, March 2002.
[33] A. Khong, P. Naylor, J. Benesty, A low delay and fast converging improved proportionate algorithm for sparse system identification, EURASIP J. Audio Speech Music Process. (2007).
[34] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Statist. Soc. Ser. B (Methodol.) 58 (1) (1996) 267–288.
[35] J. Benesty, T. Gänsler, D. Morgan, M. Sondhi, S. Gay, Advances in Network and Acoustic Echo Cancellation, Springer, 2001.
[36] R. Frenzel, Freisprechen in gestörter Umgebung, PhD thesis, Technical University of Darmstadt, Fortschritt-Berichte VDI, Reihe 10, Nr. 228, VDI Verlag, Düsseldorf, 1992.

Author's personal copy

[37] A. Mader, H. Puder, G. Schmidt, Step-size control for acoustic echo cancellation filters – an overview, Signal Process. 80 (9) (2000) 1697–1719.
[38] P. Meissner, R. Wehrmann, J. van der List, A comparative analysis of Kalman and gradient methods for adaptive echo cancellation, AEÜ, Int. J. Electron. Commun. 34 (12) (1980) 485–492.
[39] S. Yamamoto, S. Kitayama, An adaptive echo canceller with variable step gain method, Trans. IECE Jpn. E65 (1) (1982) 1–8.
[40] B. Nitsch, A frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency-domain, Signal Process. 80 (9) (2000) 1733–1745.
[41] J. Benesty, D. Morgan, J. Cho, A new class of double talk detectors based on cross-correlation, IEEE Trans. Speech Audio Process. 8 (2000) 168–172.
[42] D. Duttweiler, A twelve-channel digital echo canceler, IEEE Trans. Commun. 26 (1978) 647–653.
[43] K. Ochiai, T. Araseki, T. Ogihara, Echo canceler with two echo path models, IEEE Trans. Commun. 25 (1977) 589–595.
[44] H. Buchner, W. Kellermann, Improved Kalman gain computation for multichannel frequency-domain adaptive filtering and application to acoustic echo cancellation, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, May 2002, pp. 1909–1912.
[45] V. Myllylä, G. Schmidt, Pseudo-optimal regularization for affine projection algorithms, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, May 2002, pp. 1917–1920.
[46] E. Hänsler, G. Schmidt, Control of LMS-type adaptive filters, in: S. Haykin, B. Widrow (Eds.), Least-Mean-Square Adaptive Filters, Wiley, 2003, pp. 175–240.
[47] G. Enzner, R. Martin, P. Vary, On spectral estimation of residual echo in hands-free telephony, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Darmstadt (Germany), September 2001, pp. 211–214.
[48] G. Enzner, R. Martin, P. Vary, Partitioned residual echo power estimation for frequency-domain acoustic echo cancellation and postfiltering, Eur. Trans. Telecommun. 13 (2) (2002) 103–114.
[49] G. Enzner, R. Martin, P. Vary, Unbiased residual echo power estimation for hands-free telephony, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando (Florida), May 2002, pp. 1869–1893.
[50] S. Gustafsson, Enhancement of Audio Signals by Combined Acoustic Echo Cancellation and Noise Reduction, PhD thesis, RWTH Aachen, in: Peter Vary (Ed.), Aachener Beiträge zu digitalen Nachrichtensystemen, Band 11, Verlag der Augustinus Buchhandlung, Aachen, June 1999.
[51] P. Huber, Robust Statistics, Wiley, New York, 1981.
[52] H. Buchner, J. Benesty, T. Gänsler, W. Kellermann, Robust extended multidelay filter and double-talk detector for acoustic echo cancellation, IEEE Trans. Speech Audio Process. 14 (9) (2006).
[53] T. Gänsler, S. Gay, M. Sondhi, J. Benesty, Double-talk robust fast converging algorithms for network echo cancellation, IEEE Trans. Speech Audio Process. 8 (2000) 656–663.
[54] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, New York, 2001.
[55] H. Buchner, Broadband Adaptive MIMO Filter Theory, Springer-Verlag, Berlin, 2012.
[56] H. Buchner, W. Kellermann, A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation, in: Proceedings of Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA), Trento, Italy, May 2008.
[57] H. Buchner, R. Aichner, W. Kellermann, TRINICON: a versatile framework for multichannel blind signal processing, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 3, Montreal, Canada, May 2004, pp. 889–892.
[58] S. Gustafsson, R. Martin, P. Vary, Combined acoustic echo control and noise reduction for hands-free telephony, Signal Process. 64 (1) (1998) 21–32.

Author's personal copy

[59] R. Martin, Freisprecheinrichtungen mit mehrkanaliger Echokompensation und Störgeräuschreduktion, PhD thesis, RWTH Aachen, in: Peter Vary (Ed.), Aachener Beiträge zu digitalen Nachrichtensystemen, Band 3, Verlag der Augustinus Buchhandlung, Aachen, June 1995.
[60] R. Martin, J. Altenhöner, Coupled adaptive filters for acoustic echo control and noise reduction, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 1995, pp. 3043–3046.
[61] R. Martin, S. Gustafsson, The echo shaping approach to acoustic echo control, Speech Commun. 20 (3–4) (1996) 181–190.
[62] R. Le Bouquin Jeannès, P. Scalart, G. Faucon, C. Beaugeant, Combined noise and echo reduction in hands-free systems: a survey, IEEE Trans. Speech Audio Process. (2001) 808–820.
[63] G. Enzner, P. Vary, Robust and elegant, purely statistical adaptation of acoustic echo canceler and postfilter, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Kyoto (Japan), September 2003, pp. 43–46.
[64] E. Hänsler, G. Schmidt, Hands-free telephones – joint control of echo cancellation and postfiltering, Signal Process. 80 (11) (2000) 2295–2305.
[65] G. Enzner, M. Pauls, VDA Bewertung der modellbasierten Optimalfilterung für die Echounterdrückung in KFZ-Freisprechsystemen, in: Proceedings of Deutsche Jahrestagung für Akustik (DAGA), Dresden (Germany), March 2008.
[66] P. Eneroth, S. Gay, T. Gänsler, J. Benesty, A real-time stereophonic acoustic subband echo canceler, in: S. Gay, J. Benesty (Eds.), Acoustic Signal Processing for Telecommunications, Kluwer Academic Publishers, 2000, pp. 135–152.
[67] W. Pauler, Akustik-Test: Kfz-Freisprecheinrichtungen, Funkschau 11 (2000) 20–27.
[68] "test"-Magazine, Testbericht Freisprechanlagen: "Wer billig kauft…", Stiftung Warentest, February 2002, pp. 21–25.
[69] H.W. Gierlich, F. Kettler, Speech quality – a multidimensional problem: an approach to combine different quality parameters, in: Proceedings of Congrès Français d'Acoustique (CFA), Deutsche Jahrestagung für Akustik (DAGA), Strasbourg, France, March 2004.
[70] R. Gray, A. Buzo, A. Gray, Y. Matsuyama, Distortion measures for speech processing, IEEE Trans. Acoust. Speech Signal Process. 28 (4) (1980) 367–376.
[71] S. Quackenbush, T. Barnwell, M. Clements, Objectives Measures of Speech Quality, Prentice-Hall, Englewood Cliffs, New Jersey, 1988.
[72] ITU-T Rec. G.107, The E-model, a computational model for use in transmission planning, March 2003.
[73] ITU-T Rec. P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, February 2001.
[74] ITU-T Rec. P.862.1, Mapping function for transforming P.862 raw result scores to MOS-LQO, November 2003.
[75] ITU-T Rec. G.164, Echo suppressors, November 1988.
[76] ITU-T Rec. G.165, Echo cancellers, March 1993.
[77] ITU-T Rec. P.340, Transmission characteristics and speech quality parameters of hands-free terminals, May 2000.
[78] ITU-T Rec. P.342, Transmission characteristics for telephone band (300–3400 Hz) digital loudspeaking and hands-free telephony terminals, May 2000.
[79] ITU-T Rec. P.800, Methods for subjective determination of transmission quality, August 1996.
[80] ITU-T Rec. P.800.1, Mean opinion score (MOS) terminology, March 2003.
[81] ITU-T Rec. P.832, Subjective performance evaluation of hands-free terminals, May 2000.
[82] ITU-T Rec. P.835, Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, November 2003.

Author's personal copy

[83] F. Kettler, Kfz-Freisprecheinrichtungen: Test-Spezifikation für den Guten Ton, Funkschau 24 (2002) 50–53.
[84] Verband der Automobilindustrie, VDA Specification for Car Hands-free Terminals (Version 1.4), Verband der Automobilindustrie, December 2002.
[85] H.W. Gierlich, F. Kettler, Advanced speech quality testing of modern telecommunication equipment: an overview, Signal Process. 86 (6) (2006) 1327–1340 (special issue on applied speech and audio processing).
[86] S. Möller, F. Kettler, H.W. Gierlich, S. Poschen, N. Cote, A. Raake, M. Wältermann, Extending the E-model for capturing noise reduction and echo canceller impairments, J. Audio Eng. Soc. 60 (3) (2012) 165–174.
[87] L. Nunes, F. Avila, A. Tygel, L. Biscainho, B. Lee, A. Said, R. Schafer, A parametric objective quality assessment tool for speech signals degraded by acoustic echo, IEEE Trans. Audio Speech Lang. Process. 20 (8) (2012) 2181–2190.
[88] G. Enzner, P. Vary, New insights into the statistical signal model and the performance bounds of acoustic echo control, in: Proceedings of European Signal Processing Conference (EUSIPCO), Antalya, Turkey, September 2005.
[89] G.H. Golub, C.F. van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore, 1996.
[90] J.G. Proakis, D.G. Manolakis, Digital Signal Processing: Principles, Algorithms, and Applications, Prentice-Hall, Upper Saddle River, New Jersey, 1996.
[91] L.L. Scharf, Statistical Signal Processing, Addison-Wesley Publishing Company, 1991.
[92] R. Unbehauen, Systemtheorie 1, seventh ed., R. Oldenbourg Verlag, München Wien, 1997.
[93] R. Kalman, A new approach to linear filtering and prediction problems, Trans. ASME J. Basic Eng. 82 (1960) 35–45.
[94] G. Enzner, Bayesian inference model for applications of time-varying acoustic system identification, in: Proceedings of European Signal Processing Conference (EUSIPCO), Aalborg, Denmark, August 2010.
[95] D. Lippuner, Model-Based Step-Size Control for Adaptive Filters, PhD thesis, ETH Zürich (Diss. No. 14461), in: Hans-Andrea Loeliger (Ed.), Series in Signal and Information Processing, vol. 8, Hartung-Gorre Verlag, Konstanz, January 2002.
[96] D. Lippuner, A.N. Kälin, Tracking behavior of model-based adaptive FIR filters with noise variance estimation, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Pocono Manor, Pennsylvania, September 1999, pp. 156–159.
[97] S. Malik, G. Enzner, Model-based vs. traditional frequency-domain adaptive filtering in the presence of continuous double-talk and acoustic echo path variability, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Seattle, WA, USA, September 2008.
[98] S. Malik, G. Enzner, Online maximum-likelihood learning of time-varying dynamical models in block-frequency domain, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, TX, USA, March 2010.
[99] M.M. Sondhi, D.A. Berkeley, Silencing echoes on the telephony network, Proc. IEEE (1980) 948–963.
[100] C. Faller, J. Chen, Suppressing acoustic echo in a sampled auditory envelope space, IEEE Trans. Acoust. Speech Signal Process. (2005) 1048–1062.
[101] C. Faller, C. Tournery, Estimating the delay and coloration effect of the acoustic echo path for low complexity echo suppression, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), September 2005.
[102] A.N. Birkett, R.A. Goubran, Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects, in: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 1995, pp. 13–16.
[103] V. Turbin, A. Gilloire, P. Scalart, Enhancement of speech corrupted by musical noise, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1997, pp. 307–310.

Author's personal copy

[104] A. Favrot, C. Faller, M. Kallinger, F. Kuech, M. Schmidt, Acoustic echo control based on temporal fluctuations of short-time spectra, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), September 2008.
[105] E. Robledo-Arnuncio, T.S. Wada, B.-H. Juang, On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation, in: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2007, pp. 34–37.
[106] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process. (1979) 113–120.
[107] M. Azaria, D. Hertz, Time delay estimation by generalized cross correlation methods, IEEE Trans. Acoust. Speech Signal Process. (1984) 3689–3692.
[108] W. Etter, G.S. Moschytz, Noise reduction by noise-adaptive spectral magnitude expansion, J. Audio Eng. Soc. (1994) 341–349.
[109] M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by musical noise, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1979, pp. 208–211.
[110] F. Kuech, M. Kallinger, M. Schmidt, C. Faller, A. Favrot, Acoustic echo suppression based on separation of stationary and non-stationary echo components, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), September 2008.
[111] C. Faller, F. Baumgarte, Binaural cue coding – Part II: Schemes and applications, IEEE Trans. Speech Audio Process. 11 (6) (2003).
[112] B.R. Glasberg, B.C.J. Moore, Derivation of auditory filter shapes from notched-noise datas, Hear. Res. (1990) 103–138.
[113] C. Faller, Perceptually motivated low complexity acoustic echo control, in: Preprint 114th Convention of the Audio Engineering Society, March 2003.
[114] A. Papoulis, S.U. Pillai, Probability, Random Variables and Stochastic Processes, McGraw-Hill, New York, 2002.
[115] J. Benesty, D. Morgan, M. Sondhi, A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation, IEEE Trans. Speech Audio Process. 6 (2) (1998) 156–165.
[116] T. Gänsler, J. Benesty, Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview, Int. J. Adapt. Control Signal Process. (2000).
[117] S. Shimauchi, S. Makino, Stereo projection echo canceller with true echo path estimation, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Detroit, MI, USA, May 1995, pp. 3059–3062.
[118] A. Sugiyama, Y. Joncour, A. Hirano, A stereo echo canceler with correct echo-path identification on an input-sliding technique, IEEE Trans. Signal Process. 49 (11) (2001) 2577–2587.
[119] H. Buchner, W. Kellermann, Acoustic echo cancellation for two and more reproduction channels, in: Conference Rec. IEEE International Workshop on Acoustic Echo and Noise Control (IWAENC), Darmstadt, Germany, September 2001, pp. 99–102.
[120] J. Herre, H. Buchner, W. Kellermann, Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement, in: Proceedigns of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, HI, USA, April 2007.
[121] M. Sondhi, D. Morgan, Stereophonic acoustic echo cancellation – an overview of the fundamental problem, IEEE Signal Process. Lett. 2 (8) (1995) 148–151.
[122] H. Buchner, R. Aichner, W. Kellermann, Blind source separation for convolutive mixtures exploiting non-gaussianity, nonwhiteness, and nonstationarity, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Kyoto, Japan, September 2003, pp. 223–226.

Author's personal copy

[123] H. Buchner, R. Aichner, W. Kellermann, Blind source separation for convolutive mixtures: a unified treatment, in: J. Benesty, Y. Huang (Eds.), Audio Signal Processing for Next-Generation Multimedia Communication Systems, Kluwer Academic Publishers, Boston, April 2004, pp. 255–293.

[124] H. Buchner, R. Aichner, W. Kellermann, Relation between blind system identification and convolutive blind source separation, in: Proceedings of Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA), Piscataway, NJ, USA, March 2005.

[125] H. Buchner, R. Aichner, W. Kellermann, TRINICON-based blind system identification with application to multiple-source localization and separation, in: S. Makino, T.-W. Lee, S. Sawada (Eds.), Blind Speech Separation, Springer, Berlin, September 2007, pp. 101–147.

[126] H. Buchner, Acoustic echo cancellation for multiple reproduction channels: from first principles to real-time solutions, in: Proceedings of ITG Conference on Speech Communication, ITG Report No. 211, Aachen, Germany, October 2008.

[127] D. Morgan, J. Hall, J. Benesty, Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation, IEEE Trans. Speech Audio Process. 9 (5) (2001) 686–696.

[128] T. Gänsler, P. Eneroth, Influence of audio coding on stereophonic acoustic echo cancellation, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 1998, pp. 3649–3652.

[129] M. Ali, Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, WA, USA, May 1998, pp. 3689–3692.

[130] T. Hoya, J. Chambers, P. Naylor, Low complexity of ϵ-NLMS algorithms and subband structures for stereophonic acoustic echo cancellation, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Pocono Manor, NJ, USA, September 1999.

[131] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, MIT Press, Cambridge, MA, 1997.

[132] H. Malvar, A modulated complex lapped transform and its application to audio processing, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Phoenix, AZ, USA, 1999, pp. 1421–1424.

[133] A. Berkhout, D. de Vries, P. Vogel, Acoustic control by wave field synthesis, J. Acoust. Soc. Am. 93 (5) (1993) 2764–2778.

[134] H. Buchner, S. Spors, W. Herbordt, W. Kellermann, Wave-domain adaptive filtering for acoustic human-machine interfaces based on wavefield analysis and synthesis, in: Proceedings of European Signal Processing Conference (EUSIPCO), Vienna, Austria, September 2004.

[135] S. Spors, H. Buchner, R. Rabenstein, W. Herbordt, Active listening room compensation for massive multichannel sound reproduction systems using wave-domain adaptive filtering, J. Acoust. Soc. Am. 122 (1) (2007) 354–369.

[136] H. Buchner, S. Spors, A general derivation of wave-domain adaptive filtering and application to acoustic echo cancellation, in: Proceedings of Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, October 2008.

[137] K. Helwani, H. Buchner, S. Spors, Source-domain adaptive filtering for MIMO systems with application to acoustic echo cancellation, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, TX, USA, March 2010.

[138] O. Hoshuyama, A. Sugiyama, An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, May 2006.

[139] B.S. Nollett, D.L. Jones, Nonlinear echo cancellation for hands-free speakerphones, in: Proceedings of IEEE Workshop on Nonlinear Signal and Image Processing (NSIP), Michigan, September 1997.

Author's personal copy

[140] A. Stenger, W. Kellermann, Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling, Signal Process. 80 (2000) 1741–1760.
[141] F. Kuech, A. Mitnacht, W. Kellermann, Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, March 2005.
[142] A. Stenger, R. Rabenstein, Adaptive Volterra filters for acoustic echo cancellation, in: Proceedings of IEEE Workshop on Nonlinear Signal and Image Processing (NSIP), Antalya, June 1999.
[143] F. Kuech, W. Kellermann, Nonlinear acoustic echo cancellation, in: E. Hänsler, G. Schmidt (Eds.), Topics in Acoustic Echo and Noise Control, Springer, Berlin, 2006.
[144] W. Klippel, Dynamic measurement and interpretation of the nonlinear parameters of electrodynamic loudspeakers, J. Audio Eng. Soc. 38 (12) (1990) 944–955.
[145] H. Schurer, Linearization of Electroacoustic Transducers, Enschede: Print Partners Ipskamp, 1997.
[146] S. Malik, G. Enzner, Variational Bayesian inference for nonlinear acoustic echo cancellation using adaptive cascade modeling, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, March 2012.
[147] F. Kuech, W. Kellermann, Orthogonalized power filters for nonlinear acoustic echo cancellation, Signal Process. 86 (2006) 1168–1181.
[148] S. Malik, G. Enzner, State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation, IEEE Trans. Audio Speech Language Process. 20 (7) (2012) 2065–2079.
[149] A. Guérin, G. Faucon, R.L. Bouquin-Jeannès, Nonlinear acoustic echo cancellation based on Volterra filters, IEEE Trans. Acoust. Speech Signal Process. 11 (6) (2003) 672–683.
[150] F. Kuech, W. Kellermann, A novel multidelay adaptive algorithm for Volterra filters in diagonal coordinate representation, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, May 2004.
[151] M. Zeller, W. Kellermann, Fast and robust adaptation of DFT-domain Volterra filters in diagonal coordinates using iterated coefficient updates, IEEE Trans. Signal Process. 58 (3) (2010) 1589–1604.
[152] M. Zeller, L.A. Azpicueta-Ruiz, J. Arenas-Garcia, W. Kellermann, Adaptive Volterra filters with evolutionary quadratic kernels using a combination scheme for memory control, IEEE Trans. Signal Process. 59 (4) (2011) 1449–1464.
[153] J. Sotschek, Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die deutsche Sprache, in: Proceedings of Deutsche Jahrestagung für Akustik (DAGA), 1984, pp. 873–876.
[154] R. Streicher, F.A. Everest, The New Stereo Soundbook, second ed., Audio Engineering Associates, Pasadena, CA, USA, 1998.
[155] G. Enzner, P. Vary, On the problem of acoustic echo control in cellular networks, in: Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Eindhoven, The Netherlands, September 2005, pp. 213–216.
[156] M. Pawig, P. Vary, Energy efficiency of network-based acoustic echo control in mobile radio, in: 10. ITG Conference on Speech Communication, Braunschweig, Germany, September 2012.
[157] K. Helwani, H. Buchner, S. Spors, Spatio-temporal signal preprocessing for multichannel acoustic echo cancellation, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, CZ, May 2011.
[158] S. Malik, G. Enzner, Fourier expansion of Hammerstein models for nonlinear acoustic system identification, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, CZ, May 2011.