3

PROBLEM FORMULATIONS IN ACOUSTIC MODELLING

Contents

IMMERSIVE speech communications often take place in multisource reverberant environments where interfering signals may deteriorate the speech intelligibility. In order to tackle such limitations, IAIs aims at modelling the acoustic channel by means of adaptive filtering algorithms. In this chapter we introduce a set of problems which limit the achievable communication quality, and how to address these problems using adaptive filtering algorithms. Moreover, we briefly describe some of the main acoustic applications in which it is possible to employ IAIs based on adaptive algorithms.

3.1 MAIN CHARACTERISTICS OF ACOUSTIC CHANNELS

The problems to address in the modelling of acoustic channels are substantially different from those occurring in other communication channels, such as wireless or fibre channels. This is due to the fact that acoustic channels possess distinctive characteristics that set them apart from other kinds of transmission channels and focus attention on the development of more effective algorithms for IAIs. In the following we summarize some of the main characteristics of acoustic channels that must be taken into account in designing adaptive algorithms for IAIs.

3.1.1 Linearity and shift-invariance

An acoustic channel can be definitely labelled as a *linear shift-invariant* (LSI) system [65]. Linearity and shift-invariance are the two most important properties for simplifying the analysis and design of discrete-time system and often such characteristics do not belong to other communication channels. A linear system ought to satisfy the rules of *homogeneity* and *additivity* which are

the basis of the *principle of superposition*. For a homogeneous system, scaling the input by a constant results in the output being scaled by the same constant. For an additive system, the response of the system to a sum of two signals is the sum of the two responses. A system is shift-invariant when a time shift in its input leads to the same shift in its output. Therefore, taking into account these properties, an LSI system can be easily characterized by its impulse response. Once the impulse response is known, it is possible to foresee the response of the LSI system to any possible input stimuli.

3.1.2 Modelling by FIR filters

The AIR is usually very long. However, *finite impulse response* (FIR) filters are more frequently used than *infinite impulse response* (IIR) filters in acoustic applications. This choice is justified by the fact that the stability of FIR filters is easily controllable; moreover, there are a large number of adaptive algorithms providing good performance for FIR filters, thus allowing an accurate modelling of the acoustic channel [65, 120].

3.1.3 Time-varying AIR

Like many other communication channels with different physical medium, acoustic channels are inherently *time-varying* systems. In immersive speech communications sound sources are free to move in the environment. Moreover, even a change of atmospheric conditions in the environment may cause a variation of the AIR. However, this time-varying property usually does not prevent the use of FIR filters to model acoustic channels since acoustic systems generally change slowly compared to the length of their AIR [65]. Therefore, dividing time into periods, it is possible to assume that in each period the acoustic channel is stationary and can be modelled by means of an FIR filter.

3.1.4 Frequency selectivity

Acoustic waves are pressure disturbances propagating in the air. With spherical radiation and spreading, the inverse-square law rules and the sound level falls off as a function of distance from the sound source. As a rule of thumb, the loss is 6 dB for every doubling of distance. But when acoustic sound propagates over a long distance (usually greater than 30 m), an excess attenuation of the high-frequency components can often be observed in addition to the normal inverse-square losses, which indicates that the acoustical channel is *frequency selective* [65]. The level of this high-frequency excess attenuation is highly dependent on the air humidity and other atmospheric conditions.

The inverse-square law governs free-space propagation of sound. But in such enclosures as offices, conference rooms, and cars, acoustic waveforms might be reflected many times by the enclosure surfaces before they reach a microphone. The attenuation to the reflection is generally frequency-dependent. However, for audio signals this dependency is usually not significant, unlike radio-frequency signals in indoor wireless communication. For acoustic channels in these environments, it is the aspect of multipath propagation that leads to frequency-selective characteristics. Frequency-selective fading is viewed in the frequency domain. In the time domain, it is called *multipath delay spread* and induces sound reverberation analogous to inter-symbol interference observed in data communications.

3.1.5 Reverberation time

Room reverberation is usually regarded as destructive since sound in reverberant environments is subject to temporal and spectral smearing, which results in distortion in both the envelope and fine structure of the acoustic signal [65]. If the sound is speech, then speech intelligibility will be impaired. However, room reverberation is not always detrimental. Although it may not be realized consciously, reverberation is one of many cues used by a

listener for sound source localization and orientation in a given space. In addition, reverberation adds "warmth" to sound due to the colorization effect, which is very important to musical quality. The balance between sound clarity and spaciousness is the key to the design of attractive acoustic spaces and audio instruments, while the balance is achieved controlling the level of reverberation.

The level of reverberation is typically measured by the *reverberation time*, T_{60}, which was introduced by Sabine [118] and is now a part of the ISO (*International Organization for Standardization*) reverberation measurement procedure. The reverberation time is defined as the length of time that it takes the reverberation to decay 60 dB from the level of the original sound. The most widely used method for measuring the sound decay curves is to employ an excitation signal and record the acoustic channel's response with a microphone.

3.1.6 Channel invertibility and minimum-phase

The *invertibility* of an acoustic channel is of particular interest in many acoustic applications such as speech enhancement and dereverberation. A system is invertible if the input to the system can be uniquely determined by processing the output with a stable filter [65]. In other words, there exists a stable inverse filter that exactly compensates the effect of the invertible system. A stable, causal, rational system requires that its poles be inside the unit circle. Therefore, a stable, causal system has a stable and causal inverse only if both its poles and zeros are inside the unit circle. Such a system is commonly referred to as a *minimum-phase* system [65].

Unfortunately AIRs are almost never minimum-phase [94]. This implies that perfect deconvolution of an acoustic channel can be accomplished only with an "acausal" filter. This may not be a serious problem for off-line processing since we can incorporate an overall time delay in the inverse filter and make it causal. But the delay is usually quite long for acoustic channels and the idea is difficult to implement with real-time systems.

3.1.7 Multichannel diversity

In *multiple-input multiple-output* (MIMO) systems, one of the most important feature is the *channel diversity*, which implies that different channels of a MIMO system would have no modes in common [65]. If the channels are modelled as FIR filters, channel diversity means that their transfer functions share no common zeros, or in other words, they are co-prime polynomials.

However, in this dissertation we deal with adaptive algorithms for *single-input single-output* (SISO) systems; therefore, for possible future extension of such algorithms in the multichannel domain, the characteristic of multichannel diversity will have to be take into account.

3.1.8 Sparse acoustic impulse response

Recently, it has been recognized that most AIRs are sparse in their nature, i.e., only a small percentage of the impulse response components have a significant magnitude while the rest are zero or small [40]. This characteristic can be exploited by a class of adaptive algorithms, named *proportionate adaptive filters* [40, 13, 100], in order to improve their performance in terms of initial convergence and tracking. Proportionate adaptive algorithms will be extensively discuss in Part II of this dissertation.

3.2 LIMITATIONS AND PROBLEMS IN ACOUSTIC PATH MODELLING

As previously said, adaptive filtering algorithms in IAIs aim at modelling an acoustic channel through the estimate of the AIR generated by the acoustic coupling between a loudspeaker and a microphone. However, the AIR estimate becomes more critical when the acoustic path is affected by adverse conditions of the environment. The design of an adaptive algorithm has to take into account such problems in order to provide anyway an accurate estimate of the AIR that allows to preserve the quality of an immersive speech

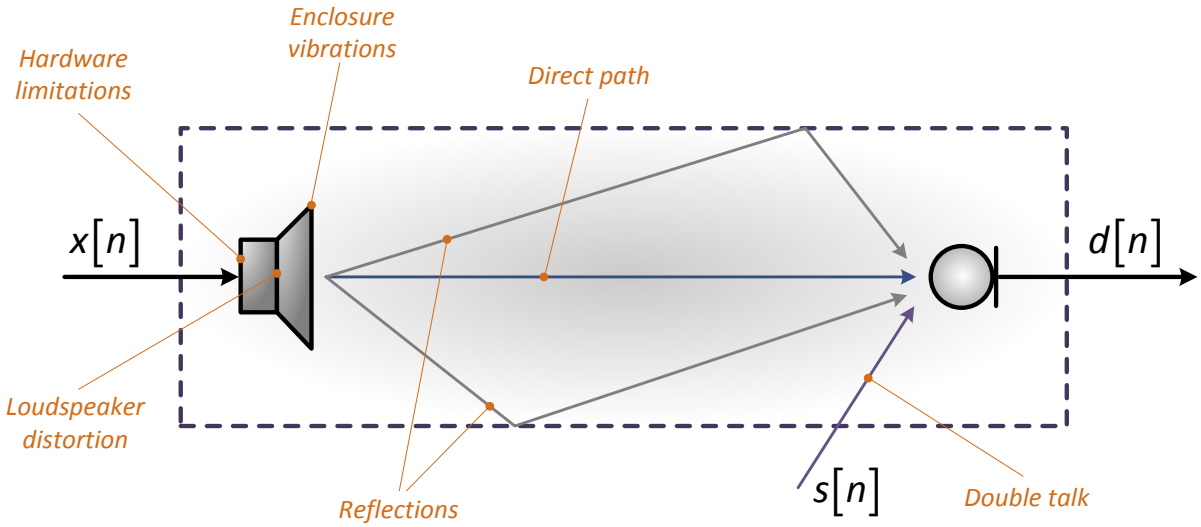

Fig. 3.1: *An acoustic interface.*

communication.

In this section we introduce a brief overview on such problems which limit the performance of an AIR modelling; they may be essentially labelled as linear or nonlinear events and are depicted in Fig. 3.1.

3.2.1 Linear limitations

Hardware limitations

Hardware limitations include thermal and impulsive circuit noise from amplifiers, and DSP related noise such as truncation, finite word lengths and characteristics of the particular algorithm being used [18]. These limitations are often caused by low-quality electronic components used in low-cost acoustic interfaces. This kind of problem essentially affects the step size value of the adaptive algorithm which may need to be very small, thus leading to a decrease of convergence performance at steady-state. Therefore, this limitation requires a good trade-off between convergence rate and precision.

Under-modelling of the AIR

As said in Par. 3.1.2, the modelling of the AIR is usually performed by means of FIR filters. However, this entails some difficulties in designing the filter, and the first and foremost one is the choice of the filter length. Indeed, it is very difficult to *a priori* know the *exact* length of the AIR, and, anyway, it usually requires a large number of filter coefficients, that is unpracticable for a real-time implementation. This is the reason why the habit is to choose a filter length smaller than the actual length of the AIR, thus leading to an *under-modelling* of the AIR. The remaining unmodelled tail portion of the AIR manifests itself as a finite error at the output of the processor. However, blindly increasing the number of taps results in added complexity, greater algorithmic noise and slower convergence. Therefore, this limitation requires a proper setting of the step size value in order to avoid this further error contribution at the output of the modelling system.

Nonstationary environment

The initial convergence of a particular algorithm identifies the room configuration, however as objects move and the input characteristics become nonstationary, the tracking ability of the algorithm becomes important. For example, although Hessian-based algorithms, such as the *recursive least squares* (RLS) algorithm, have fast convergence, it has been found that algorithms based on instantaneous gradient estimates, like the *normalized least mean square* (NLMS), actually outperform Hessian-based algorithms when nonstationarities occur [120, 18].

Double talk

The *double talk* event occurs when an interfering speech signal is present and is superimposed over the acoustic path to model. In order to solve this problem a *double talk detector* (DTD) is usually adopted [57], which stops the filter adaptation in presence of double talk in order to preserve the desired

speech. A DTD is a good mean to meet the contradictory requirement of low divergence rate and fast convergence in acoustic channel modelling. However, not ever a DTD provides desired performance, since an optimal DTD is difficult to realize and may be even very expensive from a computational point of view.

3.2.2 Nonlinear limitations

Loudspeaker distortions

Generated mainly in the loudspeaker, *nonlinear distortions* effectively put a limit on the achievable quality of algorithms based on linear mechanics [147, 18]. In addition to the direct loudspeaker effects, secondary nonlinear effects such as *rattling* can be considered nonlinear in nature. Rattling is very difficult, if not impossible to model. However, the loudspeaker nonlinearity is weak and may therefore be modelled accurately with nonlinear models. Loudspeaker distortions represent a very difficult problem to solve since they may be highly time-varying, thus leading to a kind of nonlinearity with memory.

Enclosure vibrations

A major part of the AIR is due to loudspeaker/microphone/enclosure coupling which is stationary in nature and larger in amplitude than a speech signal. The particular adaptive algorithm used will devote a portion of its computation to adapt these AIR coefficients which may be better modelled by another method. *Whistling* can occur in small orifices in sealed enclosures. This whistling is essentially chaotic in nature and can be a problem if it occurs close to the microphone [18]. Such vibrations, especially in the lower voice frequencies, causes significant nonlinearities which may seriously impair the intelligibility of a hands-free speech communication.

3.3 ACOUSTIC ECHO CANCELLATION

A typical application of acoustic channel modelling is definitely the *acoustic echo cancellation* (AEC). Acoustic echo in a hands-free voice communication system is produced by the acoustic coupling between a loudspeaker and a microphone, as depicted in Fig. 3.2. The perception of an echo depends on not only its level but also its delay [66]. Through long-distance transmission, the echo features a long delay time and would significantly reduce the quality of voice communication. When the delay approaches a quarter of a second, most people find it difficult to carry on a normal conversation. Full-duplex voice telecommunication was implausible, if not impossible, before the echo cancellation theory was developed by Bell Labs researchers in the 1960s [132]. For an immersive audio system with several microphones and loudspeakers, multiple echo paths need to be identified. Regardless of how many microphones there are, AEC is always carried out individually with respect to each of them. But the number of loudspeakers present in the system draws a theoretical difference between monophonic (one loudspeaker) and multichannel (multiple loudspeakers) echo cancellations in the difficulty of tracking the echo paths [66].

In echo cancellation, the source (loudspeaker) signals are known. So echo control is theoretically a well-posed problem [66], and its practical applications have been relatively more successful than the control of the other types of noise (such as additive noise, reverberation and unwanted speech) in which blind or semiblind methods have to be incorporated.

Historically, the study of acoustic echo cancellation substantially enriched the adaptive filtering and system identification literature. Indeed, an adaptive filter plays a central role in a monophonic echo cancellation system. It attempts to dynamically identify the acoustic echo path. As long as the channel impulse response of the echo path can be quickly and accurately determined, it is then straightforward to generate a good estimate of the echo and subtract it from the microphone signal. Since the loudspeaker signal as the reference is available,

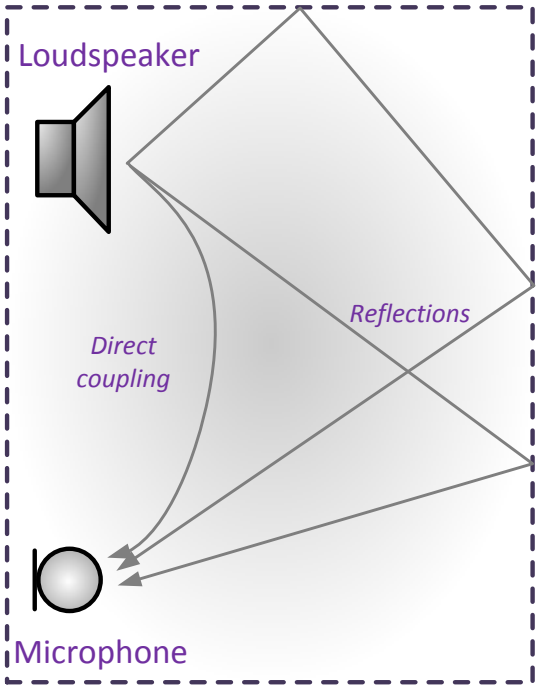

Fig. 3.2: *Microphone-loudspeaker acoustic coupling.*

numerous nonblind adaptive filtering methods for system identification are applicable for solving this problem [131, 12, 66].

In order to better comprehend AEC application, let us introduce a brief description of the processing performed by an *acoustic echo canceller* in the context of a teleconferencing communication between two (or more) users located in different environments. As it is possible to notice from the scheme in Fig. 3.3, at n-th time instant, the speech signal coming from the remote user, also known as *far-end*, and denoted as $x\,[n]$, arrives at the other side of communication and is reproduced by the loudspeaker. During the reproduction the far-end signal may result distorted by loudspeaker nonlinearities. Moreover, being the speech communication *immersive*, the far-end signal reproduced by the loudspeaker is acquired by the microphone(s) of the acoustic interface used by the local user, or also said *near-end*. The acoustic coupling between the microphone and the loudspeaker is characterized by an acoustic path which contains information about the environment reverberations. The signal

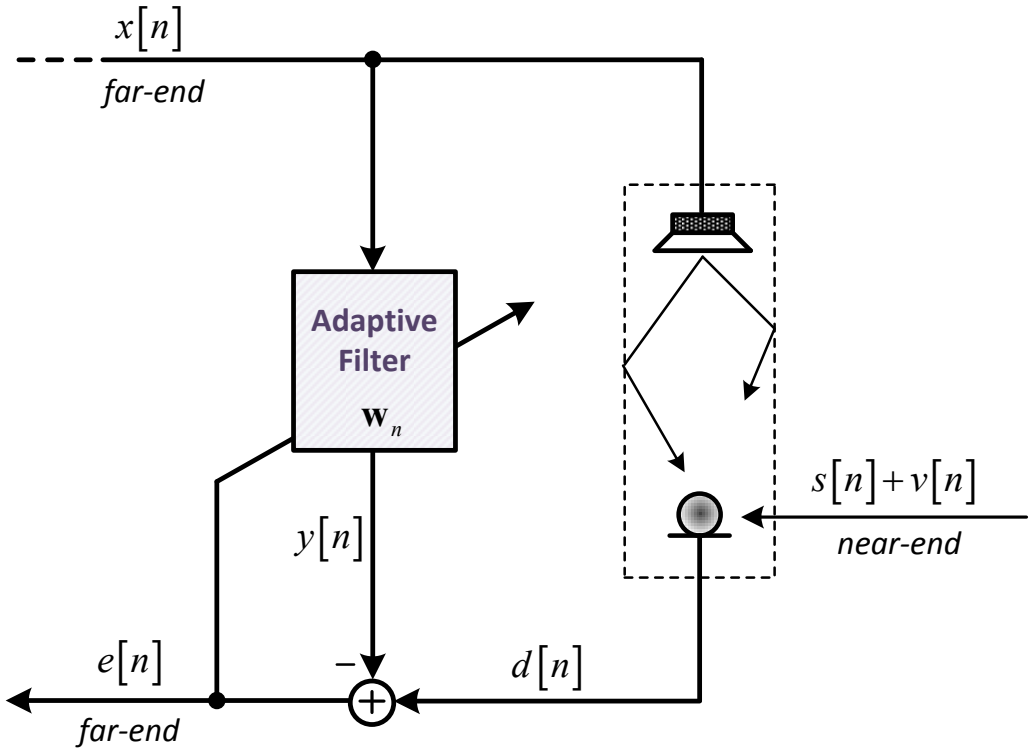

Fig. 3.3: *Processing scheme of an acoustic echo canceller.*

emitted by the loudspeaker and acquired by the microphone represents the *echo signal*, which may be possibly superimposed on the near-end contribution that is the desired information for the far-end user. The near-end signal is composed of the near-end speech signal $s\,[n]$ with the addition of background noise $v\,[n]$. In literature, the overall microphone signal is usually named as *desired signal* and it is denoted with $d\,[n]$. At the same time, the far-end signal $x\,[n]$ is processed by the acoustic echo canceller in order to estimate the AIR between microphone and loudspeaker. The output signal of this filtering process, $y\,[n]$, represents the estimated echo signal which is then subtracted by the microphone signal $d\,[n]$, preserving the near-end information, to the end of generating the *error signal* $e\,[n]$ that is sent to the far-end user.

AEC represents an exhaustive application in hands-free speech commu-

nications since it includes a set of problems common to the whole sector of acoustic scene analysis: the estimate of the impulse response, the presence of nonstationary elements in the environment, the presence of unwanted interfering signals, the presence of nonlinearities [12]. Moreover, AEC allows to obtain a complete evaluation of the adaptive filtering algorithms that may be used afterwards also in other acoustic applications, such as adaptive beamforming, noise reduction, speech dereverberation, speech enhancement, etc.

3.4 PERFORMANCE MEASURE

In order to evaluate performance of adaptive filtering algorithm in AEC applications two measures are usually computed: the echo return loss enhancement and the normalized misalignment.

3.4.1 Echo return loss enhancement

The *echo return loss enhancement* (ERLE) is defined by G.168 as "the attenuation of the echo signal as it passes through the send path of an echo canceller". The ERLE results from the ratio in dB between the instantaneous power of the desired signal $d\,[n]$, i.e. the microphone signal, and the instantaneous power of the residual echo signal $e\,[n]$ [57]:

$$\mathrm{ERLE}\,[n] = 10 \log \frac{\mathrm{E}\left\{d^2\,[n]\right\}}{\mathrm{E}\left\{e^2\,[n]\right\}} \tag{3.1}$$

A large value of the ERLE denotes a good performance of the acoustic echo canceller, while a small value of the ERLE denotes a significant presence of the echo signal in the processed signal.

In Fig. 3.4 the limitation effects on the maximum achievable ERLE is represented. It is possible to see that a first important limit is posed by the acoustic environment due above all to reflections and nonstationary signals. However, more important limits are generated by the presence of nonlinearities in the

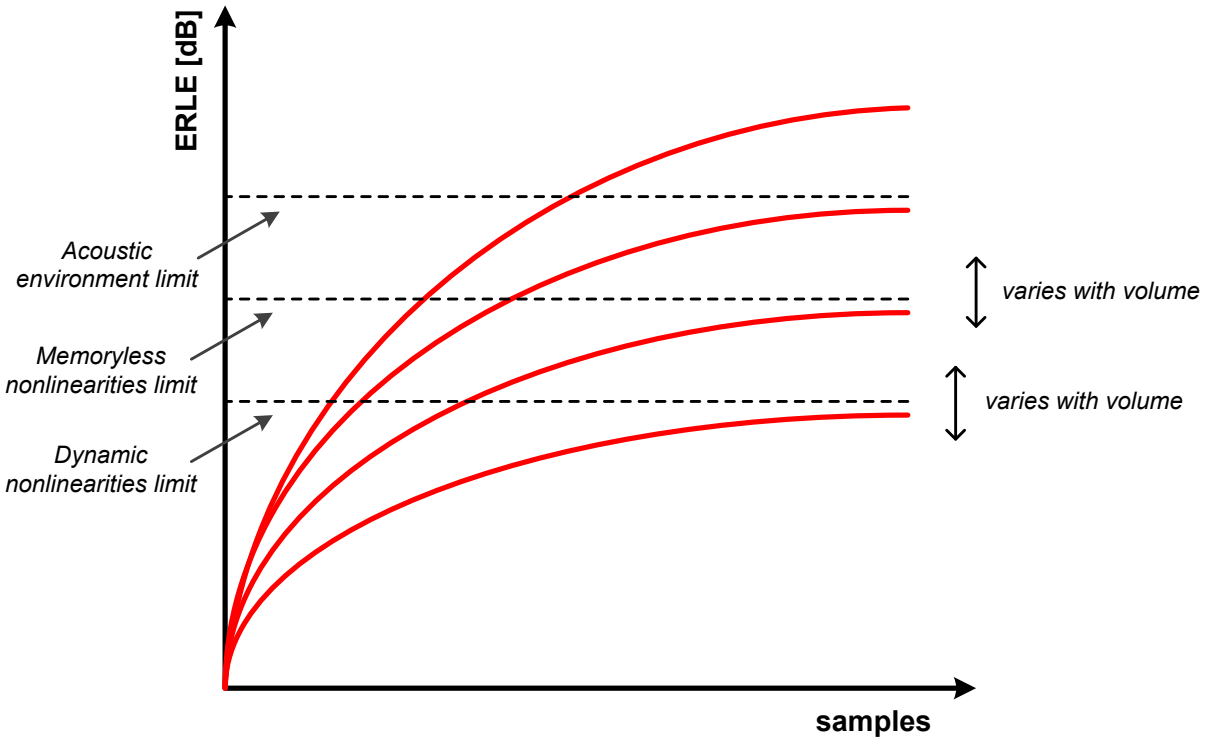

Fig. 3.4: *Limitation effects on the achievable ERLE.*

echo path, and in particular by nonlinearities with memory, i.e. those nonlinearities which are originated by dynamic systems. These limits posed by nonlinearities also depends on volume and frequency variations and may be particularly harmful to speech quality when *intermodulation distortions* occur at low frequencies.

As it is possible to notice from equation (3.1), the ERLE is a measure that depends on the minimization of the error signal. This allows to use the ERLE in the evaluation of both linear and nonlinear echo cancellers. However, the ERLE does not highlight sufficiently small variations of the adaptive algorithm; moreover, a large value of the ERLE does not guarantee as much large degree of speech quality. Due to these reasons, according to our opinion, the ERLE is not always the best performance measure to adopt in order to evaluate an adaptive filter in AEC applications; however, in literature the ERLE remains the most used performance measure to evaluate echo cancellers.

3.4.2 Normalized misalignment

Another important performance measure is the *normalized misalignment* which quantifies how "well" an adaptive filter converges to the impulse response of the system that needs to be identified [12]. It is defined in dB as:

$$\mathcal{M} = 20 \log_{10} \left(\frac{\left\| \mathbf{w}^{\mathrm{opt}} - \hat{\mathbf{w}}_n \right\|_2}{\left\| \mathbf{w}^{\mathrm{opt}} \right\|_2} \right) \tag{3.2}$$

where $\mathbf{w}^{\mathrm{opt}}$ is the optimal solution to estimate, i.e. the AIR, and $\hat{\mathbf{w}}_n$ is the filter estimate by the adaptive filter.

Unlike the ERLE, the normalized misalignment depends on the coefficients of the adaptive filter instead of the error signal, thus leading to some advantages and drawbacks. The most significant drawback is the fact that the normalized misalignment cannot be used to evaluate adaptive filters in presence of nonlinearities. This is due to the fact that nonlinearities are not taken into account in the optimal solution while they affect the filter estimate, thus the normalized misalignment does not have sense in this case. However, the normalized misalignment, unlike the ERLE, allows to have a complete evaluation of a linear adaptive algorithm in terms of convergence rate, tracking, and accuracy of the solution at steady-state. Moreover, the behaviour of the normalized misalignment also reflects the perceived quality of the processed speech signal. In fact, when the normalized misalignment shows a jumpy behaviour usually the processed signal may display some musical noise.

Such analysis focus the attention on the evaluation of the performance of adaptive filters in the nonlinear case, in which it is not possible to exploit a such important measure as the normalized misalignment. This might be definitely matter of future research.