ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0081

EDITE - ED 130

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Christelle YEMDJI TCHASSI

le 18 Juin 2013

Acoustic echo cancellation for single- and dual-microphone devices

Application to mobile phones

Directeur de thèse : **Nicholas EVANS**

Jury

M. Raymond KNOPP, Professeur, EURECOM, Sophia-Antipolis	President du Jury
M. Patrick NAYLOR, Professeur, Imperial College, London	Rapporteur
M. Marc MOONEN, Professeur, Katholieke Universiteit, Leuven	Rapporteur
M. Christophe BEAUGEANT, Docteur, Intel Mobile communications, Sophia-Antipolis	Examinateur
M. Alexandre GUERIN, Docteur, Orange Labs, Rennes	Examinateur
M. Ludovick LEPAULOUX, Docteur, Intel Mobile communications, Sophia-Antipolis	Examinateur
M. Peter VARY, Professeur, Institut für Nachrichtengeräte und Datenverarbeitung, Aachen	Examinateur

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

THÈSE

"Les hommes peuvent atteindre un but commun sans emprunter les memes voies."

Amadou Hampate Ba – Aspects de la civilisation africaine

"On a deux vies, et la deuxième commence le jour où l'on se rend compte qu'on en a qu'une"

Confucuis

Remerciements

Si l'on m'avait dit à mon arrivée en France en 2003 que je ferai un doctorat, j'aurai cru à une farce ;-). Tout a commencé en Octobre 2009 lorsque je me lance un peu à taton dans cette fabuleuse aventure qu'à été cette thèse. Aujourd'hui alors que j'achève mon doctorat, je n'ai pas de regret. Bien au contraire, j'éprouve une certaine satisfaction. Les travaux menés pendant cette these ont été possible grâce à l'appui et au soutien de plusieurs personnes envers lesquelles je souhaite exprimer ma reconnaissance.

J'aimerais tout d'abord remercier Christophe Beaugeant de m'avoir accorder sa confiance à l'issue de mon stage chez Infineon technologies et de l'accompagnement pendant ces 3 années qui auront été, je l'espere passionnantes pour toi aussi. Je souhaite également adresser ma gratitude a mon encadrant académique Dr. Nicholas Evans pour ton accompagnement, ta compréhension et surtout ta patience. Je pense notamment à la redaction des articles qui n'a pas toujours été de tout repos.

Ma gratitude va aussi l'egard de l'équipe de l'IND à Aachen et en particulier au Prof. Vary pour avoir accepté de m'accueillir au sein de son équipe pour 5 mois. Mon séjour au sein de cette équipe aura été le point d'orgue d'une part importante de mes travaux.

Je suis entrée dans le monde du traitement de la parole par le biais d'un stage effectué chez Orange Labs sous la responsibilité d'Alexandre Guerin que je souhaite remercier pour cette "initiation" au traitement de la parole et d'avoir accepté de faire partir de mon jury. Ma reconnaissance va aussi à Moctar Mossi pour nos innombrables discussions techniques et instructives. C'était un plaisir de partager un bureau avec toi.

Tout au long de cette thèse, j'ai egalement recu beaucoup de soutien de mon entourage. Je souhaite remercier ma famille, qui se trouve pour l'essentiel au Cameroun, pour son soutien et ses encouragements. L'echo aura souvent ete present dans nos communications mais

J'éprouve egalement une immense reconnaissance envers ma famille lilloise. En particulier aux Tsoméné de Wattignies, merci de votre soutien pendant cette periode charnière de ma vie. Il serait impossible pour moi de ne pas parler des "Franciscaines ;-) de Lille": Armelle, Irene N. , Irene K., Sarah et Yollande. Vous etes les meilleures les filles. Merci d'avoir toujours été là pour moi - bientot 10 ans que vous me supporter...

Je souhaite également faire un clin d'oeil aux copains d'Antibes en particulier Mounira et Ronald. Je terminerai en adressant un merci à Miriam pour son amitie: on les aura faits ensembles nos doctorats.

Abstract

With the increased flexibility and mobility which they provide, mobile terminals are arguably the most popular and widespread telecommunications devices of the present day. Mobile terminals are used in widely different and adverse conditions such as in handsfree mode or in noisy environments. Particularly in hands-free mode, part of the far-end voice signal from the loudspeaker is coupled to the microphone. Furthermore, in noisy environments the microphone also captures the ambient noise in addition to the useful near-end speech signal. In consequence, mobile terminals are generally equipped with speech signal processing algorithms in order to maintain acceptable speech quality. This thesis mainly focuses on echo cancellation.

Acoustic echo cancellation is generally achieved through adaptive filtering followed by echo postfiltering. Adaptive echo cancellation aims to estimate the echo signal recorded by the microphone. The echo postfilter is used to attenuate the residual echo. Echo postfiltering often operates in the subband or frequency domain. Filtering in the frequency or subband domain is appealing on a computational complexity point of view. However, subband filtering suffers from significant signal delay whilst frequency domain filtering suffers from time domain aliasing due to circular convolution. Alternative filtering methods can be use to avoid these problems. The first contribution in this thesis aims to assess different filtering methods for combined noise reduction and echo postfiltering. Assessments show that all filtering methods approximately achieve the same amount of echo suppression. However, they are not equivalent in terms of perceived speech quality: time domain filtering methods introduce some crackling while other methods introduce musical noise.

Although adaptive echo cancellation and echo postfilter both target the same problem, they are generally implemented separately. We propose a synchronized approach to adaptive echo cancellation and echo postfiltering. We also introduce a new stepsize computation methods. The synchronization approach and stepsize proposed method achieves a significant increase of the convergence rate and robustness of the adaptive filter.

Recent approaches to improve the speech quality are based on multi-microphone devices. In line with this trend, dual-microphone mobile devices can be found on the market. The last part of this thesis deals with dual-microphone echo control. Recordings with dual-microphone devices show us that a significant level difference is observed between the microphone signals. Based on this observation, we propose a frequency domain based double-talk detector. The proposed DTD is assessed with a single microphone echo postfilter. Assessment show that the use of the double-talk detector leads to increased echo suppression and slightly reduced distortion.

Our echo processing scheme is still composed of adaptive filtering followed by postfiltering. The novelty lies withtin the postfilter which uses two microphone signals instead of one. We introduced two novel approaches to echo postfiltering. The first is based on the level difference between the microphone and is suitable for a specific arrangement of the transducer on the device. The second postfilter is not restricted to a specific arrangements of the transducer. This solution is also advantageous as it can be used to tackle non-linear echo problem. It turns out that the proposed dual-microphone postfilter achieve good echo suppression whist keeping the distortion of the useful speech low.

Contents

List of Figures

List of notations

$u(n)$	Time domain signal
$u(k, i)$	FFT or subband signal of $u(n)$
$u^*(k, i)$	Complex conjugate of $u(k, i)$
$x(n)$	Loudspeaker signal
$x_{nl}(n)$	Non-linear loudspeaker signal such that $x_{nl}(n) = f(x(n))$ with $f(.)$ being a non-linear transformation
$h(n)$	Acoustic echo path impulse response
$d(n)$	Acoustic echo signal picked up by the microphone
$s(n)$	Near-end speech signal
$b(n)$	Ambiant noise
$y(n)$	Microphone signal
k	Time frame index
i	Frequency bin index
F_s	Sampling frequency
$\hat{u}(n), \hat{u}(k, i)$	Estimate of $u(n)$ and $u(k, i)$
$\Phi^{uu}(k, i)$	Power spectral density of signal $u(n)$
$\Phi^{uv}(k, i)$	Cross-power spectral density between signals $u(n)$ and $v(n)$
R	Hop size
M	FFT size or number of number of subbands such that i ranges from 0 to $M - 1$
α, β	Smoothing constants
$p(n), q(n)$	Analysis and synthesis filter or window
SER	Signal to echo ratio
SNR	Signal to noise ratio
η	Residual echo overestimation factor

List of abbreviations

AEC	Adaptive echo cancellation
ASFB	Analysis synthesis filter bank
CD	Cepstral distance
DDA	Decision directed approach
DFT	Discrete Fourier transform
DM	Dual-microphone
DT	Double-talk
DTD	Double-talk detector
ERLE	Echo return loss enhancement
FBE	Filter bank equalizer
FIR	Finite impulse response
GLC	Gain loss control
IDFT	Inverse discrete Fourier transform
LDF	Low delay filter
LEM	Loudspeaker-enclosure-microphone
OLA	Overlap add
PLD	Power level difference
PPN	Polyphase network
PSD	Power spectral density
RTF	Relative transfer function
SA	Speech attenuation
SER	Signal-to-echo ratio
SFTF	Short time Fourier transform
SM	Single-microphone
SNR	Signal-to-noise ratio
THD	Total harmonic distortions

Chapter 1

Introduction

The first telephone was invented in 1876 making long-distance communication possible [Balle, 2005]. The patent for this invention was filed by Graham Bell for the account of *Bell Telephone Company.* At the time the telephone was still a privilege and it is mostly installed very calm environments (i.e. salon, office). The user was required to use both hands to do a conversation: one to hold the loudspeaker and another to hold the microphone. Although this configuration was constraining for the user, it remains optimal on acoustical point of view: it offers a good isolation of the between the loudspeaker and microphone. In addition, communications take place in high signal-to-noise conditions resulting in good speech quality conditions.

Thanks to its patent on the telephone, *Bell Telephone Company* had the monopole on this market until 1894 [Leclerc and Carré, 1995]. As the *Bell Telephone Company* technologies became public, new actors arrived on the market of telecommunications creating a rude competition. Each competitor creating a network that functioned within a city. Reaching a phone located in a different city often meant inter-connecting different network and was not always possible. When the interconnection was possible, it was very costly for the consumers. First regulations offices such as the international telecommunications union (ITU) were created in the 1930's with the aim of structuring the telecommunications network in order to facilitate inter-connectivity between different networks and to defend the interest of the consumers towards rising prices [Curien and Gensollen, 1992].

Handsfree devices were only invented in the 1957. With handsfree devices, the user freed his hands thereby offering increased comfort and flexibility [Flichy, 2004]. In return, part of the signal from the loudspeaker is recorded by the microphone: the far-end user experiences the unpleasant effect of hearing a delayed version of their voice. The first attempts at echo control were based on analog voice controlled switches [Hänsler and Schmidt, 2004]. When the near-end speaker is talking, the signal on the loudspeaker path is completely suppressed. Therefore, when both speakers were active, both sending and receiving path signals are suppressed and no communication is possible. This is the half-duplex effect - only one person can speak at a time. Full-duplex solutions to the echo problem came in the 1970's with the development of digital circuits and the invention of adaptive filters. Adaptive filters are used to obtain a real-time estimate of the loudspeaker-enclosure-microphone (LEM) system.

Mobile devices were developed in the 1980's allowing for phone conversations to take place everywhere. The use of mobile devices in noisy environments (i.e. cafe, airport, street ...) introduces the problem of additive noise in telecommunications. The first noise reduction methods aimed to suppress noise by subtracting it from the recorded microphone signal [Boll, 1979].

The earliest mobile devices weighted about 3kg whereas nowadays weights of 100g are typical. The miniaturization of mobile phones is also a source of some other speech quality degradation. One of the most important impairments comes from the loudspeaker miniaturization which can introduce a lot of non-linearities in the LEM system. As a result, echo control solutions need to account for the presence of non-linearities.

Speech quality in mobile phones is degraded by a diversity of artifacts. Among which we can cite the problem of ambient noise and that of acoustic echo. In consequence, mobile terminals are generally equipped with speech signal processing algorithms in order to maintain acceptable speech quality [Degry and Beaugeant, 2008, Hänsler and Schmidt, 2004]. This thesis focuses on echo cancellation.

1.1 State-of-the-art approaches to speech enhancement

Most approaches to acoustic echo cancellation consist of an adaptive filter followed by an echo postfilter [Benesty et al., 2007, Hänsler and Schmidt, 2004, Martin, 1995]. The adaptive filter produces an estimate of the acoustic path [Haykin, 2002]. It is then used to estimate the echo signal at the microphone in order that it be subtracted from the up-link signal. In practice, the performance of adaptive echo cancellation (AEC) is disturbed by the presence of ambient noise and/or the near-end speech signal [Hänsler and Schmidt, 2004, Haykin, 2002]. To limit the impact of such disturbance on the AEC module, double-talk detectors (DTDs) and/or noise only detectors are often used [Huang et al., 2006]. Nevertheless, some residual echo remains at the output of the adaptive filter. Consequently, postfilters are often used to further suppress residual echo. Most echo postfilters consist of an attenuation gain applied to the error signal resulting from adaptive filtering. For better performance during double-talk periods, this attenuation can be computed in the sub-band or frequency domain [Benesty et al., 2007, Hänsler and Schmidt, 2004].

Noise reduction algorithms usually operate in the frequency or sub-band domain and are generally based on the assumption that noise is an additive and relatively stationary perturbation. Commonly used noise reduction algorithms are based on an estimate of the noise power spectral density which is used to calculate a noise reduction filter [Gustafsson et al., 2001, Hänsler and Schmidt, 2004, Martin, 2001].

As illustrated in Figure 1.1, efficient speech enhancement in telecommunications terminals is guarantied by various modules. The non-linear pre-processor aims to estimate the non-linear component of the loudspeaker signal. Adaptive echo canceling furnishes an estimate of the non-linear echo signal at the microphone while echo postfiltering aims to render residual echo inaudible. The noise reduction modules are used to process the noise for the far-end and near-end speakers. In a practise, the downlink speech signal i.e. that which has to be played by the loudspeaker contains both the far-end speech signal and the ambient noise from his environment. Even in case noise reduction has already been applied to the signal prior to its transmission, the use of the downlink noise reduction can still permits to improve the speech intelligibility for the near-end speaker. The downlink noise reduction will mostly be useful for cases where the signal to echo ratio is very low.

All these modules aim to improve speech quality but most of them shave been designed and optimized individually. Moreover, this optimization does not always exploit the full hardware capacity of the device. For example, most dual-microphone mobile devices are still equipped with single microphone echo cancellation. **The objective of this thesis is to improve echo cancellation. First, we consider single microphone echo**

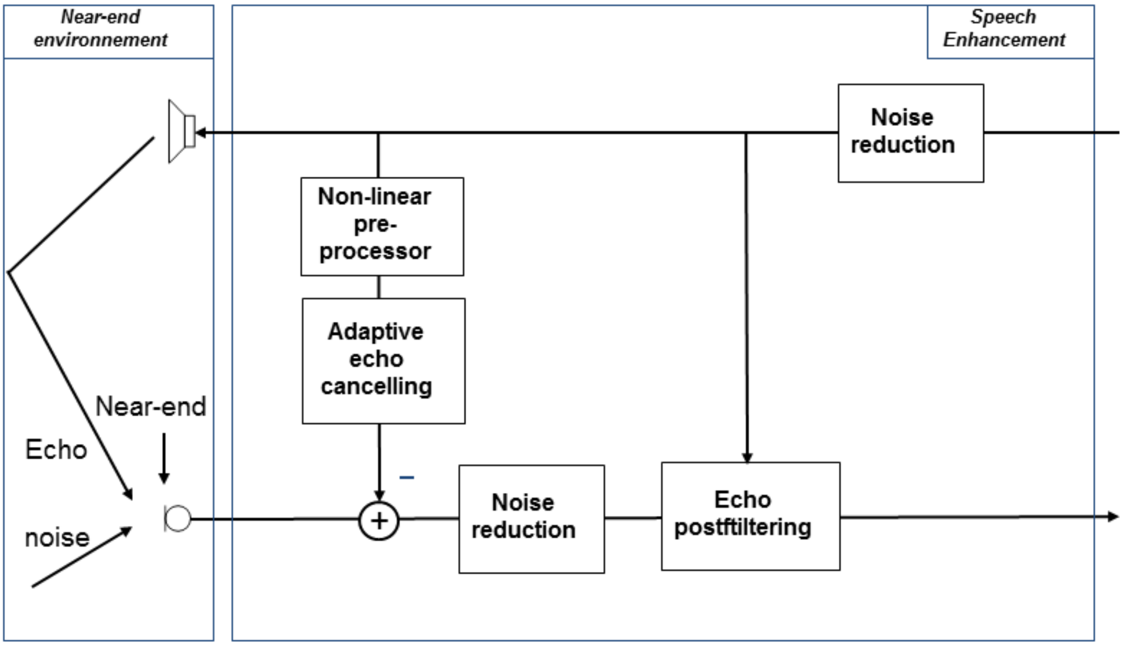


Figure 1.1: Example of speech enhancement scheme

cancellation and propose a novel architecture which accounts for the interactions between AEC and echo postfiltering. Second, we propose approaches to improve echo cancellation based on dual-microphone devices.

1.2 Contributions

In Chapter 2, we present state-of-the-art approaches to speech enhancement. Our contributions are divided in two parts addressing single and dual microphone solutions respectively.

1.2.1 Single-microphone echo cancellation

In most approaches to echo postfiltering, residual echo suppression and noise reduction (i.e. filtering of the degraded speech signal) is performed in the frequency or sub-band domain through multiplication. Frequency and sub-band domain filtering are advantageous in terms of computational simplicity [Oppenheim and Schafer, 1999] but also have some drawbacks.

Sub-band domain filtering introduces a significant delay in the output signal [Löllmann and Vary, 2007]. Delay reduction can be achieved by filtering the residual echo in the time domain. In this case, the sub-band attenuation gains are used to determine a broadband finite impulse response filter. Popular approaches to its calculation include the filter bank equalizer, the low delay filter or the inverse discrete Fourier approach [Löllmann and Vary, 2007, Steinert et al., 2008, **Yemdji** et al., 2010a]. Contributions in this thesis include a comparative assessment of alternative filtering methods to sub-band echo postfiltering [**Yemdji** et al., 2010a,b]. We showed that distortions introduced by time domain filtering methods are perceived differently from those introduced by subband filtering.

Frequency domain filtering through bin-by-bin multiplication is equivalent to circular convolution and suffers from time domain aliasing [Oppenheim and Schafer, 1999]. Distortions can be reduced using linear convolution [Oppenheim and Schafer, 1999] which is known to be computationally prohibitive for mobile terminals. In [**Yemdji** et al., 2011] we propose a low computational complexity implementation of linear convolution. A distinct advantage of the proposed implementation relates to its scalability which we exploit to manage computational complexity with only moderate degradation in speech quality. Our findings regarding subband domain and frequency domain related filtering methods are presented in Chapter 3.

The work then turns to the joint-control of AEC and echo postfiltering algorithms. The AEC and the postfilter both aim to suppress acoustic echo. Historically, each module was designed as an independent module [Hänsler and Schmidt, 2004]. Recently, however, echo control systems with synchronized adaptive echo cancellation and echo postfiltering have been investigated and have shown improved performance [Enzner and Vary, 2003, 2006, Steinert et al., 2007]. Synchronized echo control systems use the system distance (i.e. the error between the acoustic path and its estimate) to link the two modules which are in this case designed to operate in the same frequency or sub-band domain. We propose a cross domain approach to synchronized AEC and echo postfiltering and show that the proposed approach outperforms existing state-of-the-art approaches [**Yemdji** et al., 2012b]. Based on the work in [Mossi et al., 2010], our synchronization approach is advantageous for its robustness against loudspeaker non-linearities in comparison to existing synchronized systems. Our approach to synchronization is presented and assessed in Chapter 4.

1.2.2 Dual-microphone echo cancellation

The performance of single microphone echo cancellation algorithms is still limited, especially if we consider adverse situations such as handsfree configurations for which the signal-to-echo-ratio at the microphone can be low. Although recent approaches to improve speech enhancement consist in the use of multi-microphone terminals [Gannot et al., 2001, Jeub et al., 2011, Kellermann, 1997, Reuven et al., 2007a], dual microphone echo processing has not received much attention in the literature [Guo et al., 2011, Jeannes et al., 2001].

We report a study of the echo problem for dual microphone devices based on measurements with mock-up and real mobile phones. Our study is based on both handset and handsfree scenarios. Unlike existing multi-microphone echo control approaches based on beamforming [Gannot et al., 2001, Guo et al., 2011, Kellermann, 1997, Reuven et al., 2007a], we show how dual microphone echo control can be achieved through an adaptive filter followed by a postfilter [**Yemdji** et al., 2012a,c]. Our contributions regarding dual microphone echo control are two-fold.

The first part of our contribution uses the level difference between microphone signals. The reported level difference is observed for certain transducers configurations namely their arrangement on the device. The level difference is exploited to introduce a new approach to double-talk detection and a new echo suppression gain rule. The proposed double-talk detector (DTD) is appealing for its simplicity and flexibility. It can operate either in the frequency or sub-band domain as well as in the fullband domain. The level difference gain rule does not require an explicit estimate of the residual echo power spectral density. Experiments show that both the DTD and the new gain rule lead to improved echo cancellation performance compared to single microphone approaches. The proposed level difference DTD and gain rule are presented in Chapter 5.

The second part of our contribution regarding dual-microphone echo processing involves a general approach to echo postfiltering. In contrast to the level difference approaches, the proposed postfilter is not constrained to specific transducers configurations. This approach simply exploits the correlation between the two microphone signals to estimate the residual echo power spectral density [**Yemdji** et al., 2012a,c]. The proposed power spectral density estimate is readily extended to deal with loudspeaker nonlinearities. Experiments show that this method leads to more accurate power spectral density estimation and achieves better echo suppression compared to single microphone approaches as we showed in [**Yemdji** et al., 2012c]. This approach to dual-microphone echo postfiltering is presented in Chapter 6.

Publications

Part of the contributions presented in this thesis have been published by the author at international conferences: [**Yemdji** et al., 2010a,b, 2011, 2012b,c]. Most of the contributions regarding dual-microphone devices have been patented [**Yemdji** et al., 2012a]. Throughout this thesis, publications of the author are indicated in bold.

Chapter 2

Acoustic echo control in telecommunications terminals

Contents

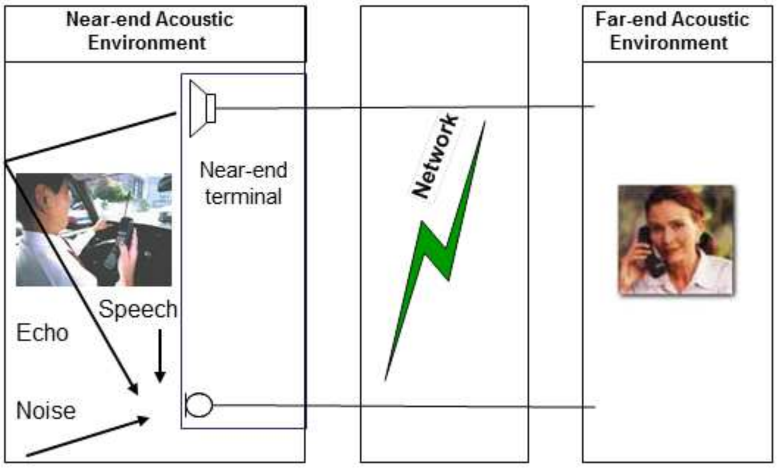


Figure 2.1: Illustration of the echo problem

The development of telecommunications is particularly marked by the advent of services such as mobile telephony or video conferencing. All these mutations aim to improve the quality of communications: comfort (hands free devices), friendliness (video calls, conversations in groups) and security (handsfree devices in cars).

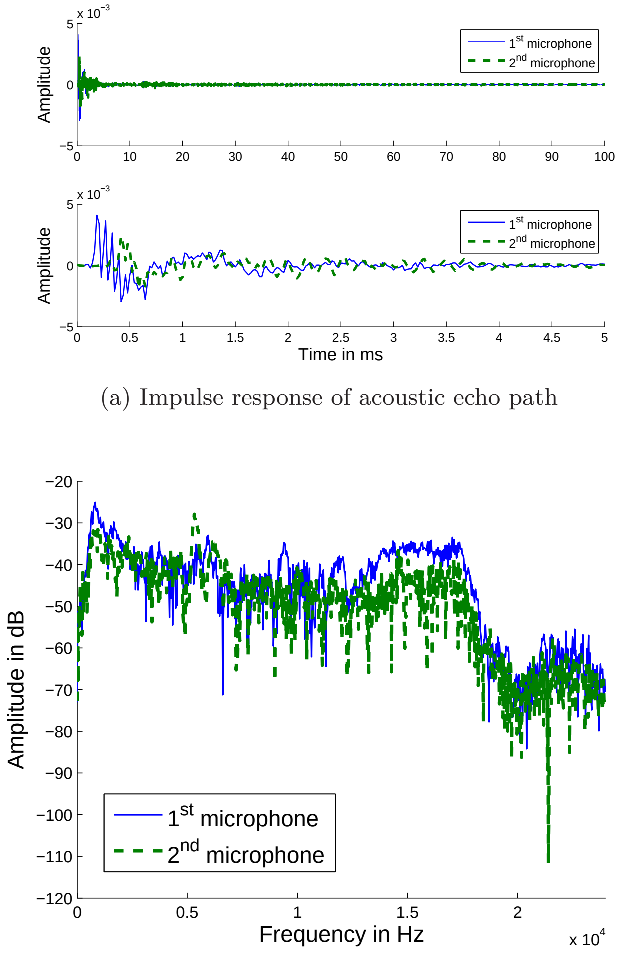
Besides the comfort provided to the user, the use of phone is some environments might degrade the speech quality. Public places such as stations, airports, etc.. are particularly noisy places. The ambient noise is picked up by the microphone just as the useful speech. With handsfree devices, part of the signal played by the loudspeaker is recorded by microphone. Because of this feedback, the speaker hears his voice with a delay (delay introduced by the transmission chain). As a result, the microphone signal contains the speech signal but also useful acoustic echo and noise: these effects can be annoying for the far-end speaker.

Speech quality is a very important aspect in telecommunications as regulation institutions such as the ITU-T or 3GPP have set some specifications regarding the echo and noise problem and quality requirements for the transmitted speech signals. Actors of the telecommunications market must develop speech enhancement algorithms to meet the specifications recommendations. Acoustic echo cancellation and noise reduction are used to tackle these problems. Speech enhancement is a topic of interest for a variety of actors such as phone designers, automotive constructor or laptop manufacturer.

This chapter deals about state-of-the-art speech enhancement algorithms and is organized as follows. In Section 2.1, we introduce the problem of acoustic echo and noise. Section 2.2 reports some state-of-the-art approaches to echo cancellation and noise reduction. Section 2.3 deals about assement that can be used to assess the performance of speech enhancement algorithms. Our conclusions are presented in 2.4.

2.1 Sound recording in telecommunications termninals

In this section, we present the problem due to acoustic echo and noise.



(a) Impulse response of acoustic echo path

(b) Frequency response of acoustic echo path

Figure 2.2: Example of measured acoustic echo path

2.1.1 Acoustic echo

In a phone conversation, the voice signal is transmitted through a communication network to a device equipped with at least one loudspeaker and one microphone. The loudspeaker plays the sound from the far-end speaker to the near-end speaker while the microphone records the voice of the near-end speaker. The voice of the near-end speaker is then transmitted to far-end speaker. But in some cases, part of sound emitted by the loudspeaker propagates in the near-end environment and is coupled the microphone of the device. As a result, the far-end speaker does not only receive the voice of the near-end speaker but also receives a delayed version of his voice: this effect is referred to as acoustic echo. As illustrated in Figure 2.1, this coupling is composed of the direct path and of the reflected paths between the loudspeaker and the microphone.

2.1.1.1 Linear echo

The coupling between the transducers of the device, also referred to as the echo path can be modeled by a finite impulse response filter. The echo signal can then be written as

$$d(n) = h(n) \star x(n) \tag{2.1}$$

where $h(n)$ represents the impulse of the echo path and $x(n)$ represents the loudspeaker signal. Figure 2.2 shows an example of echo paths measured with a mock-up handsfree mobile phone in an office environment. The mock-up phone used consists simply of a plastic box equipped with a loudspeaker and two microphones. One can refer to Annex 5.A for details about the description of the design of the mock-up phone. The use of a mock-up phone instead of a real one permits to focus only on the acoustic interactions that occur in a device. The microphones are placed such that one of them is slightly closer to the loudspeaker than the other. We denote $h_1(n)$ and $h_2(n)$ the impulse response between the loudspeaker and the first and second microphones respectively.

We see from Figure 2.2 (a) the main delay (first peak) is not the same for each microphone. The echo path is composed of the direct path and indirect paths (reflections) between the loudspeaker and the microphone of interest. The main delay for each impulse response is related to the direct path (i.e. distance) between the loudspeaker and the microphone considered [Kuttruff, 2000]. The closer the microphone is from the loudspeaker, the shorter the direct path. In our case, the main delays are of 0.2ms and 0.4ms for the first and second microphone respectively. It is also of interest to note that the amplitude of this first peak is different for each microphone. This is due to the fact that the amplitude of a propagating acoustic wave is inversely proportional to the distance between its source and the point at which it is measured. In our case, the closer the microphone is from the microphone, the higher the amplitude of the main delay will be.

The peaks that follow the main one are due to the reflections of the sound from the loudspeaker in the surrounding environment. We can see from Figure 2.2 (a) that the reflections are different for each microphone. The microphones are placed at different position on the devices and do not pick up the same reflections at the same time. The sound from the loudspeaker propagates in all directions, creating an infinite number of waves. Reflections occur when the wave encounters an obstacle: part of the incident wave then continues to propagate in a different direction (that of the reflective wave) before being picked up by the microphone. The frequency responses of the measured impulse responses are showed in Figure 2.2 (b). We can see that the acoustic path impacts on the spectral components of the loudspeaker signal: all the frequency are not equally attenuated.

2.1.1.2 Mechanical coupling

The sound wave played by the loudspeaker actually results from the vibrations of the membrane of the loudspeaker, the vibrations themselves being generated by the electric wave received from the network. The microphone records sound by transforming acoustic waves into electric waves. In the case of mobile terminals, the loudspeaker and microphone are in the same enclosure. Part of the coupling between the transducer of the phone is due to the proximity between the terminal transducers.

The coupling due to the proximity of the transducers is called mechanical coupling. Figure 2.3 illustrates the mechanical coupling for a mock-up phone. The mock-up is the same used to measure the acoustic echo paths in Figure 2.3. The mechanical coupling

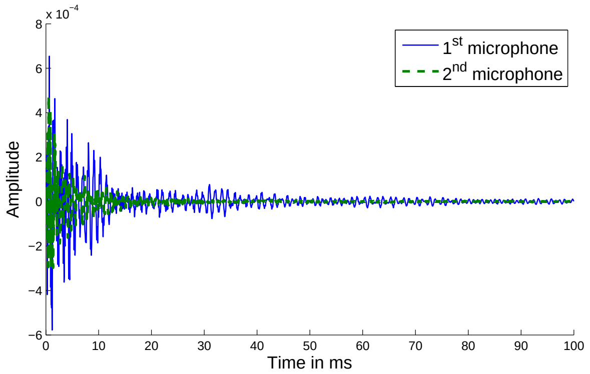
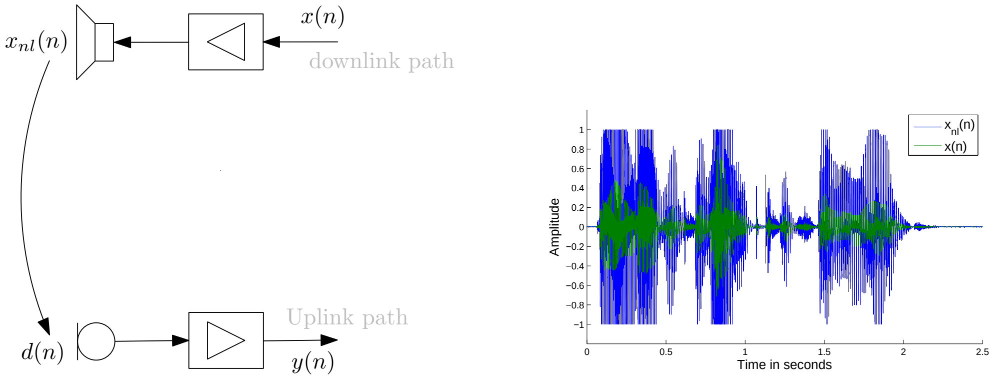


Figure 2.3: Illustration of the mechanical coupling: impulse response of mechanical coupling



(a) Scheme of a device including converter and amplifier on the sending and receiving paths

(b) Illustration of clipping for the loudspeaker signal

Figure 2.4: Effect of the amplifier used on the receiving path

is measured similarly as the acoustic echo path (of Figure 2.2) except this time, the microphones are sealed meaning the microphones should not record any signal if a sound is played on the loudspeaker. The fact that the impulse responses $h_1(n)$ and $h_2(n)$ are different from zero shows that part of the acoustic echo is due to the mechanical interaction between the transducer. The difference of amplitude between the mechanical coupling in Figure 2.3 and the acoustic coupling in Figure 2.2 shows that the effect of mechanical coupling remains marginal in comparison to the effect of the acoustic coupling.

2.1.1.3 Non-linear echo

The effects reported here only account the linear part of the acoustic echo. In reality part of acoustic echo is generated by non-linear phenomena. Non-linear acoustic echo comes from transducer saturation, digital converters and non-linearity of the loudspeaker transfer function [Guerin, 2002].

The signals transmitted via the network are digital whereas the transducers are only able to play or record analog signal. As illustrated in Figure 2.4 (a), the received digital far-end speech signal $x(n)$ is processed through an amplifier before being input to the loudspeaker itself. The amplification step permits to increase the power of the received signal

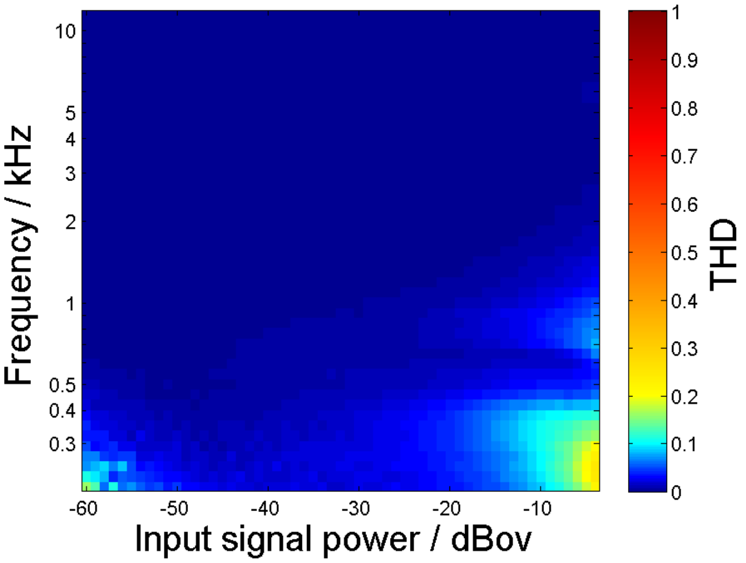


Figure 2.5: Example of total harmonic distortion for a loudspeaker. A sine with different amplitudes and frequencies is used as input signal to the loudspeaker.

such that the signal played by the loudspeaker is audible. Unfortunately, as illustrated in Figure 2.4 (b) saturations might occur during amplification. As a result, the signal at the input of the loudspeaker does not correspond to a linear transformation of the received signal $x(n)$.

In addition to non-linearities due to the loudspeaker saturations, we can mention the harmonic distortions that are due to the non-linearity of the loudspeaker transfer function. Loudspeakers are in theory designed such that their frequency response is flat for a given frequency band and input signal power. In practice, this is not the case. Figure 2.5 shows the total harmonic distortion (THD) of a loudspeaker as a function of the power and frequency of the input signal. The loudspeaker used in this distortion measurement is that of the mock-up phone used above. The THD is a measure of the amount of harmonic distortion introduced by a device. Ideally, it should be equal to zero. The signal at the input of the loudspeaker is a sine: the amplitude and frequency of the sine are used as test parameters. For a given input sine signal (i.e. with a given frequency and amplitude), we measure the power of the signal played by the loudspeaker. We see from Figure 2.5 that the loudspeaker is not linear in the low frequencies. The amount of distortions is even more important for high amplitude signals. This figure shows that high amplitude signals played by the loudspeaker will be distorted. The fact that the THD measure is not equal to zero for frequencies and amplitudes show that in practice the frequency response of a loudspeaker is not flat. The harmonics distortions of the loudspeaker can for example be modeled through a Volterra model [Birkett and Goubran, 1995a, Gao and Snelgrove, 1991]. Figure 2.5 also shows that low amplitude signal might be distorted by the loudspeaker. The distortions of the low amplitude signal are specific to the loudspeaker used and are related to its quantification limitations.

The consequence of the coupling between the loudspeaker and microphone of the phone is that the far-end hears a delayed version of his own voice. The delay due to the acoustic path is very small compared to that due to the transmission network. The level of echo depends on the acoustic path. Acoustic echo is a factor of annoyance and fatigue for the users.

In summary, in mobile phones acoustic echo is due to the coupling between the loudspeaker and the microphone. An acoustic path is characteristic of the device used and of the acoustic environment. The way the loudspeaker and microphone are placed on the device mainly defines the direct path of the acoustic echo path whereas the acoustic environment defines way the reflections occur. In the specific case of mobile phones, part of the acoustic echo is due to mechanical coupling between the transducers.

2.1.2 Ambient noise

At its early stages, telephony devices were only installed in offices and living rooms. Nowadays, phones are part of our daily lifes. Handsfree devices are used in car environments as well as in office environments. Mobile phones are used in cafes, airports. In addition to the voice of the near-end speaker and acoustic echo, part of the ambient noise is recorded by the microphone and transmitted to the far-end speaker. This is annoying for both the near-end and the far-end speakers:

- The intelligibility of the message transmitted to the far-end is degraded by the presence of noise. The presence of noise is even more annoying for the far-end speaker because he perceives the background noise from far-end speaker environment which is most likely to be different from background noise of his environment.

- The ambient noise will also overlap the signal played by the loudspeaker and therefore reducing the intelligibility for the near-end speaker.

The more the noise level increases, the more the useful signal is masked by the ambient noise resulting in less intelligibility. In other terms, the annoyance due to noise increase as the noise level increases.

2.2 Speech enhancement algorithms

We have explained how ambient noise and acoustic echo degrade speech quality in mobile terminals. Solutions to tackle these disturbances have been widely investigated in the literature. In this section, we present some state-of-the-art echo control and noise reduction algorithms. In Section 2.2.1, we present existing echo cancellation algorithms while in Section 2.2.2 we present noise reduction algorithms. Lastly, an example of speech enhancement is presented Section 2.2.3.

2.2.1 Echo processing

The first attempts to suppress acoustic echo consisted in the use of analog voice-controlled switches. With the progress of digital circuits, more efficient echo control systems have emerged. One popular tool is the adaptive filter which was back then not used because of its computational complexity. Nowadays, most echo control systems are composed of adaptive filtering followed by residual echo suppression as illustrated in Figure 2.6.

Section 2.2.1.1 deals about adaptive echo cancellation. In Section 2.2.1.2, we present existing echo postfiltering methods while Section 2.2.1.3 reports synchronized approaches to adaptive echo cancellation and echo postfiltering.

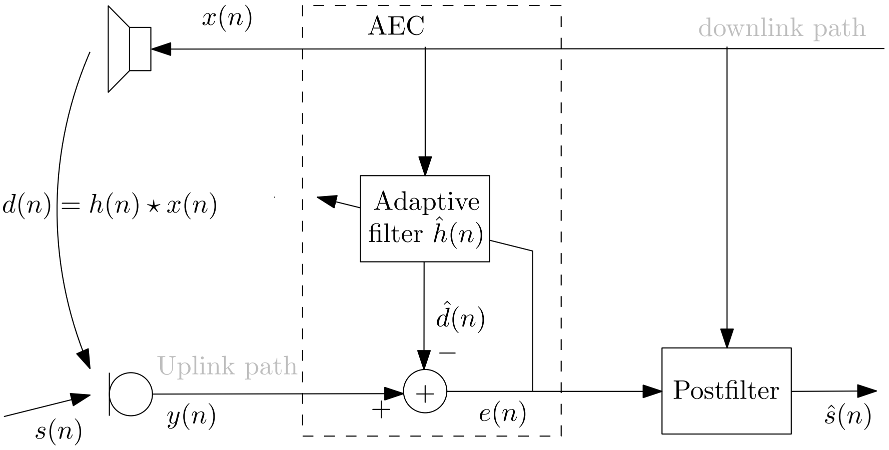


Figure 2.6: Echo control scheme

2.2.1.1 Adaptive echo cancellation

Acoustic echo results from the coupling between the loudspeaker and the microphone which can be modeled by an FIR (finite impulse response) filter. Adaptive echo cancellation aims at estimating this coupling. As shown in Figure 2.6, an adaptive filter can be used to estimate the acoustic echo path. The acoustic path estimate $\hat{h}(n)$ is then convolved with the loudspeaker signal $x(n)$ to obtain an estimate of the echo signal:

$$\hat{d}(n) = (\hat{h} \star x)(n) = \hat{\mathbf{h}}^T(n) \cdot \mathbf{x}(n). \tag{2.2}$$

where $\hat{\mathbf{h}}(n) = \begin{bmatrix} \hat{h}_0(n) & \hat{h}_1(n) & \cdots & \hat{h}_{L-1}(n) \end{bmatrix}^T$ represents the adaptive filter coefficients, $\mathbf{x}(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-L+1) \end{bmatrix}^T$ is the vector of the loudspeaker samples and L is the adaptive filter length. The update of $\hat{\mathbf{h}}(n)$ is performed by a feedback loop on the estimation error $e(n)$ proportionally to a gain denoted $\mathbf{C}(n)$:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \hat{\mathbf{C}}(n) \cdot e(n) \quad \text{with} \quad e(n) = y(n) - \hat{d}(n). \tag{2.3}$$

Equation 2.3 is the general update equation of an adaptive filter. The expression of the gain $\hat{\mathbf{C}}(n)$ depends on the minimization criteria of the adaptive algorithm (i.e. cost function) and on the assumption made on the input signals and on the acoustic path.

Steepest-descent algorithm: We denote $J(\mathbf{h}(n))$ the cost function to minimize. The steepest-descent consists in updating the adaptive filter in the direction opposite to the gradient and leads to:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \mu \frac{\partial J(\mathbf{h}(n))}{\partial \mathbf{h}(n)}. \tag{2.4}$$

where μ is the stepsize. By defining the cost function as the the mean square error (MSE)

$$J(\hat{\mathbf{h}}(n)) = E[e^2(n)], \tag{2.5}$$

where $E[\cdot]$ represent the mathematical expectation, the steepest descent algorithm becomes:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \mu E[e(n) \cdot \mathbf{x}(n)]. \tag{2.6}$$

The steepest-descent converges to the Wiener solution [Haykin, 2002]. Equation 2.6 is not use in practice because of the expectation term which cannot be computed easily.

Least mean square (LMS) algorithm: The LMS algorithm is an approximation of the steepest-descent algorithm which estimate the expectation term in Equation 2.6 by its instantaneous value. The LMS update equation is expressed as follows:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \mu e(n) \cdot \mathbf{x}(n). \tag{2.7}$$

The stepsize μ controls the convergence and stability of the LMS algorithm. To ensure convergence of the LMS, the stepsize should be such that:

$$0 < \mu < \frac{2}{\lambda_{max}} \tag{2.8}$$

where λ_{max} is the largest eigenvalue of the correlation matrix of $\mathbf{x}(n)$ [Haykin, 2002].

Normalized LMS (NLMS) algorithm: The stability of the LMS algorithm depends on the variance of the loudspeaker signal. To render the adaptive stability independent from the loudspeaker signal, the stepize is normalized by the loudspeaker signal energy. The adaptive filter then becomes

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \frac{\mu}{\mathbf{x}^T(n) \cdot \mathbf{x}(n)} \cdot \mathbf{x}(n) \cdot e(n). \tag{2.9}$$

Convergence is ensured if μ is between 0 and 2. The normalization permits to improve the convergence speed and stability of the adaptive filter. The NLMS algorithm nevertheless suffers from slow convergence especially with speech signals.

Affine projection algorithm (APA): The APA adaptive filter is also obtained by minimizing the MSE subject to a constraint on the errors over an observation window

$$\min(E[e^2(n)]) \quad \text{subject to} \quad d(n-l) = \hat{h}(n) \star x(n-l) \tag{2.10}$$

where l is an index ranging from 0 to $N-1$ and N is the length of the observation window. The APA is obtained by solving Equation 2.10 through the Lagrange method and expresses as follows:

$$\hat{h}(n+1) = \hat{h}(n) - \mu \mathbf{X}(n) (\mathbf{X}(n)\mathbf{X}(n))^{-1} \mathbf{e}(n) \tag{2.11}$$

where $\mathbf{X}(n) = \begin{bmatrix} \mathbf{x}(n) & \mathbf{x}(n-1) & \cdots & \mathbf{x}(n-N+1) \end{bmatrix}$ is a matrix of loudspeaker samples and $\mathbf{e}(n) = \begin{bmatrix} e(n) & e(n-1) & \cdots & e(n-N+1) \end{bmatrix}^T$ is the vector containing the N last error samples. Note the NLMS algorithm is a special case of the APA with $N = 1$. The APA is advantageous for its fast convergence but requires an significant computational load compared to the LMS or NLMS algorithms. For mobile devices, the computational complexity can be prohibitive factor for the choice of an algorithm.

Only the most popular approaches to adaptive echo cancellation have been presented here. Other algorithms such as LS (Least Square) or RLS (Recursive Least Square) can be used for adaptive echo canceling. One can refer to [Haykin, 2002] and [Hänsler and Schmidt, 2004] for more information.

In most systems, adaptive filters use hundreds of taps in order to produce an appropriate estimate of the acoustic path. The more taps the adaptive filter has, the better it can model the acoustic path. The longer an adaptive filter is, the more its update is computationally demanding and can be prohibitive for a real-time system. Computational complexity of adaptive filtering can be reduced through block-by-block or subband processing [Huo et al., 2001, Paleologu and Benesty, 2012, Sommen and Jayasinghe, 1988]. Block processing algorithms update the adaptive filter for a block of input samples (instead of every sample). In case of subband AEC, microphone and loudspeaker signals are split up into M subbands. Adaptive filtering in the subband domain permits to process the echo at a lower sampling rate. Subbands adaptive filters are shorter than fullband filters. In addition, the length of the subband adaptive filters can be chosen independently. Therefore instead of adapting and filtering in full-band, we have M adaptations and convolutions (in parallel) but at a lower sampling rate [Morgan and Thi, 1995].

Adaptive filter control: The convergence speed of an adaptive filter can be defined as the time this algorithm takes to reach to its optimum response or steady state. For all the algorithms mentioned above (LMS, NLMS and APA), the stepsize controls the algorithm convergence speed. Large stepsize values lead to fast convergence whereas small stepsize values lead to slow convergence [Benesty et al., 2007, chap 45]. Additionally large values of μ result in less accurate estimates of the acoustic path than small values of μ. In other terms, considering that the AEC reaches its optimum, AECs with large μ output more residual echo than AEC with small values of μ. In practice, the AEC should achieve both fast convergence and low residual echo. For this reason, time-variant stepsizes μ are often used. An optimum stepsize will be defined such that it should have a larger value during the convergence period of the AEC and smaller values after convergence periods. In practice, variable stepsizes are used to control the AEC [Benesty et al., 2006, Iqbal and Grant, 2008, Lee et al., 2009].

If we take the example of the NLMS algorithm, its optimum variable stepsize is expressed as follows:

$$\mu(n) = \frac{E[\tilde{d}^2(n)]}{E[e^2(n)]} \quad \text{with} \quad \tilde{d}(n) = d(n) - \hat{d}(n). \tag{2.12}$$

The computation of Equation 2.12 requires the knowledge of the residual echo signal $\tilde{d}(n)$ which is unknow in a real-time system.

Behavior of adaptive filters: In practice, the performance of an adaptive echo canceler is limited: the presence of ambient noise or near-end speech signal might impact the AEC. As a result, the estimation error $e(n)$ is not echo-free. The presence of residual echo in the error signal is due to the following reasons:

- The adaptive filter needs a certain time in order to converge toward its optimal response. During its converging period, the estimation error is not minimal. As a result, echo is attenuated but is still audible in the output signal $e(n)$. Further echo attenuation can be achieved by using a postfilter.

- $\hat{h}(n)$ is a FIR (Finite Impulse Response) filter of length L. To reduce computational complexity, L is smaller than the length of the real acoustic channel [Haykin, 2002]. So when the adaptive filter reaches its optimal response, the echo estimate is optimal but not equal to the echo signal $d(n)$. Again, a postfilter can be used to achieved further echo attenuation.

- Moreover, in noisy conditions (changes of the acoustic path, presence of near-end speech and noise in the near-end speaker environment) the convergence of the adaptive algorithm is disturbed. Methods to limit the impact of the noise on the AEC include the use of noise detectors and double-talk detectors (DTD) [Benesty et al., 2007, Buchner et al., 2006]. Typical use of noise detector and/or DTD consists in freezing the adaptation of the AEC during noise-only and/or double-talk (DT) periods.

- When using a mobile device, a user can hardly be static during the conversation: in some cases, the user can be walking around. Even the smallest movements of the user (i.e. nod of the head) or of the device (i.e. rotations of the mobile device) are actually seen by the AEC as changes of the acoustic path between the loudspeaker and the microphone. This means that the AEC will need some time to re-converge. Echo path change detectors are used to limit the impact of echo path on the AEC. Most approaches to such detectors are coupled with DTDs [Iqbal and Grant, 2008].

- Finally, the non-linear part of echo cannot be modeled by a FIR filter. So the AEC filter as described above cannot resolve the problem due to transducer nonlinearities [Birkett and Goubran, 1995b]. Non-linear echo cancelers are used to tackle this problem.

Double-talk detection: Most DTDs are based on a detection variable which is defined as a function of the input signals $x(n)$, $y(n)$ and $e(n)$. The detection variable is compared to a threshold and DT is eventually declared. The choice of the value of the detection threshold depends on the variable definition. A large variety of detection criteria can be found in the literature. Most features used to define the detection variable are the signals level [Duttweiler, 1978], the signals coherence [Gänsler et al., 1996] or signals correlation [Gänsler and Benesty, 2001].

Non-linear echo processing: Non-linear echo cancellation can be separated into two different categories: non-linear AEC and non-linear residual echo suppression. Non-linear AEC algorithms can be classified in two families:

- Algorithms based on loudspeaker linearization which aim at preventing the loudspeaker non-linear behavior. A pre-processor is placed on the loudspeaker path such that the signal played by the loudspeaker is a linear transformation of the received signal $x(n)$ [Furuhashi et al., 2006, Mossi et al., 2011]. The resulting acoustic path formed by the loudspeaker-enclosure-microphone system is linear. Any adaptive filter can be used to estimate the acoustic path.

- We also distinguish approaches based on the estimation of the loudspeaker nonlinearity. In this case, a pre-processor is placed before the adaptive filter and is used to estimate the non-linear loudspeaker signal $x_{nl}(n)$ [Guerin et al., 2003, Stenger and Rabenstein, 1998]. The AEC then uses this estimate of non-linear loudspeaker signal as reference signal. As a consequence the AEC is able to estimate the non-linear echo signal recorded by the microphone.

Non-linear AEC can be highly computationally demanding and suffers from slow convergence [Azpicueta-Ruiz et al., 2011, Stenger and Kellermann, 2000]. Approaches to non-linear residual echo suppression are more computationally efficient than non-linear

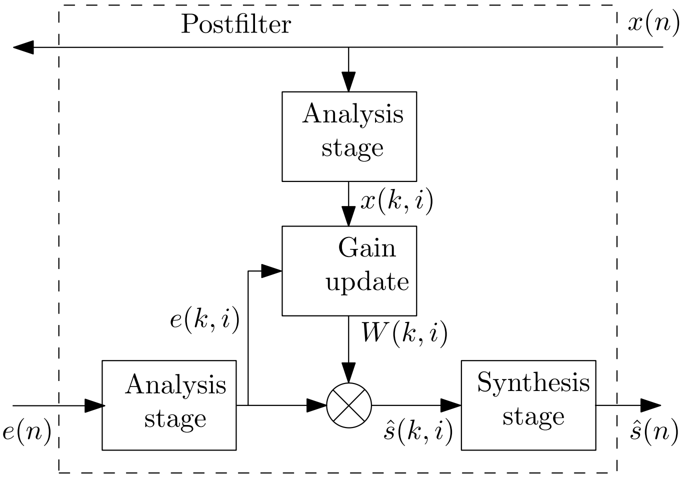


Figure 2.7: Scheme of a sub-band echo postfilter

AEC and are generally based on frequency domain postfiltering. Most approaches to non-linear AEC are based on the assumption that adaptive filter cancels all the linear echo whereas this is not the case with practical system [Hoshuyama, 2012, Shi et al., 2008].

2.2.1.2 Echo postfiltering

The objective of a residual echo suppression module is to render echo inaudible. A simple and popular approach to residual echo suppression is that of gain loss control (GLC) [Degry and Beaugeant, 2008, Hänsler and Schmidt, 2004]. GLC algorithms simply consist in applying an attenuation to the error signal. Although this gain is generally calculated as a function of the loudspeaker power [Degry and Beaugeant, 2008, Heitkamper and Walker, 1993] it impacts on near-end speech during double talk periods because it is applied independently to the presence, or not, of near-end speech.

To overcome poor double talk performance of GLC, frequency or subband echo postfilters [Beaugeant et al., 1998] are often used. Sub-band postfilters are preferred to GLC because they consist of sub-band gains and can therefore specifically target frequencies where residual echo is audible. Sub-band echo postfilters are inspired from the spectral subtraction algorithm [Boll, 1979] for noise reduction. Although sub-band echo postfilters can efficiently be used as a stand alone solution for echo canceling, it has been shown to be an efficient algorithm for residual echo suppression, i.e. in conjunction with the AEC.

Figure 2.7 illustrates the scheme of a sub-band echo postfilter. The input signals $e(n)$ and $x(n)$ are first split into frequency (or subband) domain signals $e(k,i)$ and $x(k,i)$ respectively, where i denotes the frequency and ranges from 0 to $M-1$ and k denotes the frame index. The analysis stage can be implemented through a Fourier transform or through an analysis filter bank [Allen and Rabiner, 1977, Crochiere and Rabiner, 1983]. The postfilter gains $W(k,i)$ are computed and applied to the sub-band error signal $e(k,i)$ as a multiplicative factor. The full-band near-end speech signal $\hat{s}(n)$ is recovered from the sub-band signal $\hat{s}(k,i)$ in concordance with the analysis stage.

Various gain rules for echo postfiltering can be found in the literature [Beaugeant et al., 1998, Ephraim and Malah, 1985]. The Wiener filter for echo postfiltering is derived from the minimization of the MSE in the spectral domain $E\Big[\big|s(k,i)-\hat{s}(k,i)\big|^2\Big]$ and is expressed as follows:

$$W(k,i) = \frac{\Phi^{ss}(k,i)}{\Phi^{ss}(k,i) + \eta \cdot \Phi^{\tilde{d}\tilde{d}}(k,i)} = \frac{\psi(k,i)}{1 + \eta \cdot \psi(k,i)}, \tag{2.13}$$

where η is an overestimation factor, Φ^{ss} is the near-end speech signal power spectral density, Φ^{ss} is the residual echo power spectral density and ψ is the signal to (residual) echo ratio. The overestimation factor controls the aggressiveness of the filter [Turbin et al., 1997]:

- High values of η lead to high attenuation values. During echo-only periods, the residual echo will be deleted whereas during DT periods, the near-end speech signal components will also be deleted. As a consequence, processed speech signals can exhibit musical noise during DT periods.
- In contrast, low values of η might cause insufficient echo suppression during echo-only as well as during DT periods.

There is a tradeoff to make between the amount of echo suppressed and the distortions introduced during DT periods.

Another gain rule can be derived from the minimization of the logarithm of amplitudes $E[|\log(|s(k,i)|-\log(|\hat{s}(k,i)|)|^2]$ [Ephraim and Malah, 1985]. This gain rule is often referred to as the log spectral amplitude gain rule and is expressed as:

$$W(k,i) = \frac{\psi(k,i)}{1+\psi(k,i)} \exp\Big(\frac{1}{2}\int_{\nu(k,i)}^{\infty} \frac{e^{-t}}{t} \mathrm{d}t\Big) \quad \text{with} \quad \nu(k,i) = \frac{\psi(k,i) \cdot \psi_{post}(k,i)}{1+\psi(k,i)}. \tag{2.14}$$

In [Ephraim and Malah, 1985] an approach to estimate the signal-to-echo ratio (SER) is presented as:

$$\hat{\psi}(k,i) = \beta \frac{|\hat{s}(k-1,i)|^2}{\Phi^{\tilde{d}\tilde{d}}(k-1,i)} + (1-\beta) \cdot \max\big(\frac{|e(k,i)|^2}{\Phi^{\tilde{d}\tilde{d}}(k,i)} - 1, 0\big). \tag{2.15}$$

It is of interest to note the SER estimate according to Equation 2.15 only reduces to the estimation of the residual echo PSD. All the other quantities involved in the computation of Equation 2.15 can be computed from the input signals $e(k,i)$ and $x(k,i)$. Examples of residual echo PSD estimate can be found in [Enzner et al., 2002, Habets et al., 2008, Steinert et al., 2007].

2.2.1.3 Synchronized approaches to echo cancellation

AEC and echo postfiltering both aim at suppressing the echo. At the beginning of the echo or when an echo path occurs, the AEC need some time to reach its optimum response. The level of residual echo is higher during this convergence period than in periods where the AEC has converged. The postfilter on his side is design so as to achieve more or less aggressive echo suppression. An aggressive postfilter will result in good echo suppression and strong distortions of the useful signal during double-talk periods. If the postfilter is designed so as to avoid distorting the useful signal, the residual echo might not be completely deleted. The synchronization of the AEC and echo postfilter appears as a solution to achieve a better compromise between good suppression and good DT behavior.

We can distinguish two different methods regarding synchronized echo control: systems based a static modeling of the echo path as it has been done until now and systems based on a statistical model of the echo path.

Static echo path: These approaches exploit the link between the stepsize of the AEC and the echo postfilter [Enzner and Vary, 2003, Steinert et al., 2007]. The constraint about these approaches is that the AEC and the postfilter must be in the same subband or frequency domain.

Time varying echo path: Approaches to AEC presented until now are based on the assumption that the echo path is stationary and deterministic. However, in a practical scenario, the acoustic path is time variant meaning it cannot be assumed to be stationary. These variations can sometimes be significant (e.g. : door opening or closing) or small. Small acoustic path changes can be modeled by a first-order Markov model [Enzner and Vary, 2006, Haykin, 2002].

$$h(n+1) = A \cdot h(n) + \Delta h(n) \tag{2.16}$$

where A is the transition factor which is supposed constant and comprised in between 0.99 and 0.999 [Enzner and Vary, 2006]. $\Delta h(n)$ accounts for the unpredictable changes in the acoustic path.

With this modeling of the echo path, the MMSE leads to a synchronized echo control system which is composed of a Kalman filter followed by a postfilter: both are in the frequency domain [Enzner and Vary, 2006].

$$\hat{H}(k+1,i) = A \cdot \hat{H}(k,i) + K(k,i) \cdot e(k,i) \tag{2.17}$$

where $\hat{H}(k,i)$ is the Fourier transform of the acoustic path estimate, $e(k,i)$ is the Fourier transform of the error signal $e(n)$ and $K(k,i)$ is the Kalman gain. The Kalman gain can be written in the form of $K(k,i) = \mu(k,i) \cdot X(k,i)$ where $\mu(k,i)$ stands for a variable stepsize and $X(k,i)$ is a diagonal matrix whose diagonal contains the Fourier transform of $\mathbf{x}(n)$. The computation of the Kalman gain and echo postfilter is as follows:

$$K(k,i) = \mu(k,i) \cdot X^H(k,i) = \frac{|\tilde{H}(k,i)|^2}{|\tilde{H}(k,i)|^2 \cdot |X(k,i)|^2 + \Phi^{ss}(k,i)} \cdot X^H(k,i) \tag{2.18}$$

$$W(k,i) = \frac{\Phi^{ss}(k,i)}{\Phi^{ss}(k,i) + |\tilde{H}(k,i)|^2 \cdot |X(k,i)|^2} \tag{2.19}$$

where $\tilde{H}(k,i) = H(k,i) - \hat{H}(k,i)$. The synchronization comes from the fact the Kalman gain and the postfilter both used the quantity $|\tilde{H}(k,i)|^2$. This Kalman echo control system is shown to be more robust to varying echo paths. Comparative assessments show that the Kalman AEC converges faster than standard frequency domain adaptive filtering [Malik and Enzner, 2008]. Detailed information about the validity of the acoustic path modeling and the derivation of the Kalman filter can be found in [Enzner and Vary, 2006].

2.2.2 Noise reduction

Noise reduction algorithms aim to estimate the clean speech signal. Most noise reduction algorithms operate in the frequency or subband domain and are generally based on the assumption that noise is an additive and relatively stationary perturbation. An example of noise reduction (NR) scheme is shown in Figure 2.8. The useful speech and noise signals are denoted $s(n)$ and $b(n)$. The noisy input $y(n)$ is converted from the time to the frequency (or subband) domain to obtain $y(k,i)$. The frequency domain signal $y(k,i)$ is

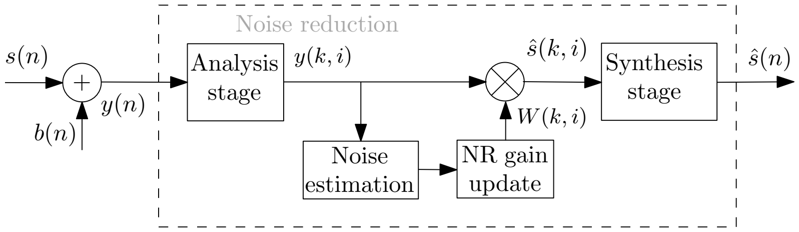


Figure 2.8: Noise reduction detailed scheme

used to estimate the noise level which is later on used to compute the attenuation gain $W(k,i)$. An estimate of the clean speech signal is obtained through multiplication of $y(k,i)$ and $W(k,i)$. Lastly, the fullband estimate of the clean speech signal is recovered from $\hat{s}(k,i)$. The efficiency of the NR module mainly depends on the choice of NR gain and noise estimate used.

2.2.2.1 Noise reduction algorithms

Spectral subtraction: Spectral subtraction is one of the first NR method [Boll, 1979]. Spectral subtraction is based on the intuitive observation that noise is an additive perturbation. Therefore, an estimate of the clean speech can be obtained through subtraction of short-term spectrum:

$$|\hat{s}(k,i)| = |y(k,i)| - \sqrt{\Phi^{bb}(k,i)} \tag{2.20}$$

where $\Phi^{bb}(k,i)$ denotes the PSD of the noise signal. Equation 2.20 can alternatively be rewritten as

$$|\hat{s}(k,i)| = \Big(1 - \frac{\sqrt{\Phi^{bb}(k,i)}}{|y(k,i)|}\Big) \cdot |y(k,i)| = W(k,i) \cdot |y(k,i)|. \tag{2.21}$$

To avoid negative attenuation gain values, the minimum value of $W(k,i)$ is set to 0.

The spectral subtraction as defined here suffers from strong distortions of the useful speech signal. More general methods for spectral subtraction can be found in [Lim and Oppenheim, 1979, Plapous, 2005, Sim et al., 1998]. These generalized spectral subtraction approaches mainly use an overestimation factor to artificially increase or decrease the noise level. High value of the overestimation lead to good NR but strong musical noise. In contrast low values of overestimation permit to reduce the musical noise at the cost of noise suppression. A compromise has to be made on the amount of noise suppression versus the amount of noise suppression [Plapous, 2005].

Wiener gain rule: The Wiener gain rule is derived from the minimization of MSE in the spectral domain $E\big[|s(k,i) - \hat{s}(k,i)|^2\big]$ and expresses as follows

$$W(k,i) = \frac{\chi(k,i)}{1+\chi(k,i)} \quad \text{with} \quad \chi(k,i) = \frac{\Phi^{ss}(k,i)}{\Phi^{bb}(k,i)} \tag{2.22}$$

where $\chi(k,i)$ is the signal to noise ratio (SNR) [Lim and Oppenheim, 1979, Scalart and Filho, 1996]. The computation of the SNR requires the knowledge of the PSD of clean and noisy signals. To overcome the fact that the PSD of the clean speech signal is unknown, most systems are based on SNR estimates.

The SNR can be approximated as:

$$\hat{\chi}(k,i) = \frac{\Phi^{yy}(k,i)}{\Phi^{bb}(k,i)} \tag{2.23}$$

such that the computation of the SNR only requires an estimation of the PSD of the noise [Cappe, 1994]. In case of high SNR (i.e. low level of noise), the SNR estimate in Equation 2.23 will be quite close to its real value. But as the noise level increases, the gap between this SNR estimate and its real value will increase. With this estimate, the phase difference between the clean speech and noise signal might also impact accuracy. The SNR estimate in Equation 2.23 is nevertheless interesting for its low computational complexity.

The SNR can alternatively be estimated as [Cappe, 1994]

$$\hat{\chi}(k,i) = \frac{\Phi^{yy}(k,i) - \Phi^{bb}(k,i)}{\Phi^{bb}(k,i)}. \tag{2.24}$$

Equation 2.24 estimates the PSD of clean speech $s(n)$ as the difference between the PSD of the noisy signal and that of the noise signal assuming the noise and clean signals always add constructively, neglecting the effects of the phase of the noise signal [Vary, 1985]. The human ear is not very sensitive to phase difference, so the phase of the noise signal can be neglected without to much distortions [Wang and Lim, 1982].

Another approach to estimation of the SNR is the decision directed approach which was introduced in [Ephraim and Malah, 1985]:

$$\hat{\chi}(k,i) = \beta \frac{|\hat{s}(k-1,i)|^2}{\Phi^{bb}(k-1,i)} + (1-\beta) \cdot \max\big(\frac{|y(k,i)|^2}{\Phi^{bb}(k,i)} - 1, 0\big) \tag{2.25}$$

where β is a smoothing constant often chosen close to 1. The decision directed approach gives a good estimate of the SNR whether high or low. Its use permits to significantly reduce the musical noise introduced by the NR module [Cappe, 1994].

2.2.2.2 Noise PSD estimation

The noise level estimate is inherent to the computation of the NR gain or SNR estimate and therefore needs to be estimated accurately. Techniques to estimate the noise include:

- The speech activity detection method: With this method, the noise estimate is updated during silence periods and frozen during speech activity periods [Gustafsson et al., 2001, Hänsler and Schmidt, 2004]. One of the main problems of this technique is the difficulty to design an efficient and robust speech activity detector.

- Some others noise PSD estimates are based on continuous tracking of the noise level. One popular estimator is the minimum statistics method. This technique exploits the fact that speech does not occupy the entire frequency band even during speech periods. The spectral and temporal holes of the speech are used to estimate the noise power by tracking the minimum noisy signal power within an observation window for a given sub-band i [Martin, 2001]. Alternative noise PSD estimates can be found in the literature [Cohen, 2003, Gerkmann and Hendriks, 2011, Goulding and Bird, 1990, Krawczyk and Gerkmann, 2012].

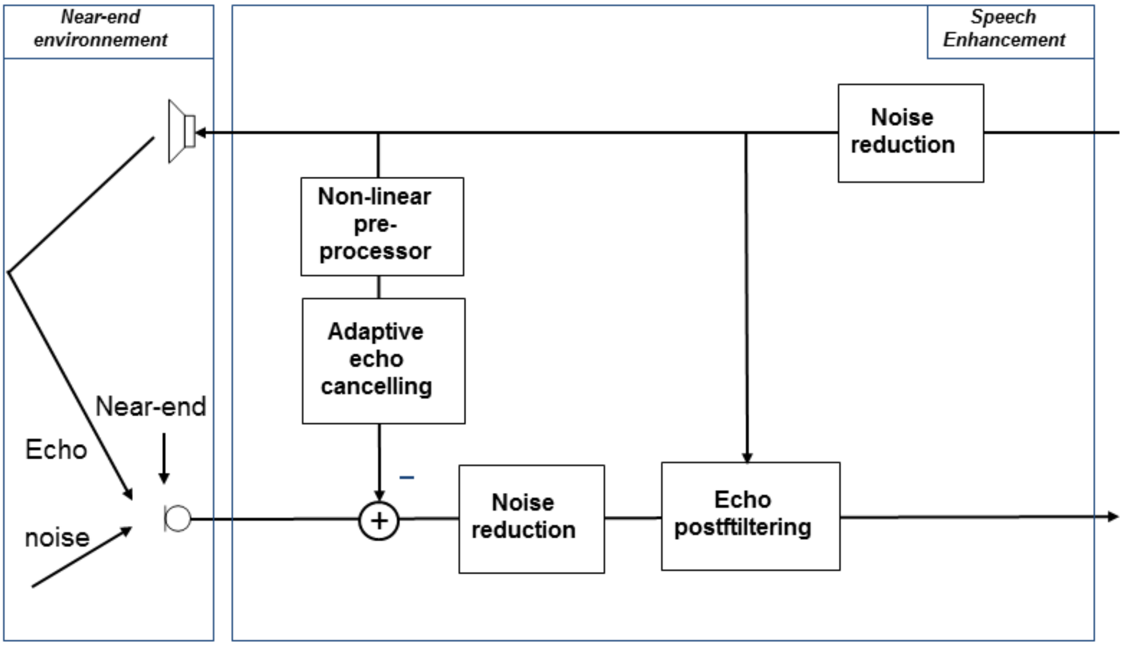


Figure 2.9: Example of speech enhancement scheme for microphone device

2.2.3 Summary of speech enhancement algorithms

In sections 2.2.1.1 and 2.2.2 we presented some state-of-the-art algorithms for echo control and noise reduction. In practice, a combination of the speech enhancement algorithms presented need to be implemented within the device. Figure 2.9 illustrates an example of speech enhancement scheme that could be used in telecommunications terminals [Degry and Beaugeant, 2008]. The non-linear preprocessor aims to estimate the non-linearities generated by the loudspeaker and is placed prior to AEC so that the AEC uses as reference signal one which is as close as possible to the one that generated the echo signal. The AEC gives an estimate of the non-linear echo. The error signal from the AEC still contains some residual echo which will be suppressed by the echo postfiltering placed after the NR module. The adaptive echo canceling aims at suppressing the non-linear echo signal whereas the postfilter processes the residual echo. The NR modules aim at reducing the noise for both speakers. NR is applied to the loudspeaker signal or downlink path in order to reduce noise that could be introduced by the network and to further attenuate noise from the far-end speaker. NR is also applied to the microphone signal or uplink path to reduce the ambient noise picked up by the near-end microphone.

The position or order of the different modules has to be chosen carefully. In [Beaugeant, 1996], it is shown that placing the NR before the AEC result in poor AEC. Indeed, applying NR to the recorded microphone does not only improve the noise but attenuation the speech signals (echo and near-end speech signals) recorded by the microphone. In that case, it then turns out that the echo signal at the input of the AEC is not necessary a linearly transformed version of the original echo signal (that recorded by the microphone). For this reason, it is preferable to place the NR after the AEC.

Figure 2.9 shows that good speech quality is achieved by numerous algorithms which each tackles a specific problem. The resulting speech enhancement scheme has a significant computational complexity and might lead to significant signal delay. Most of these algorithms have historically been designed and optimized solely. Nevertheless, as we introduced in Section 2.2.1.3, some recent approaches to improve echo suppression are based on synchronized AEC and echo postfiltering. Part of recent studies aim at reducing the computational complexity of the overall speech enhancement scheme. We can cite for example cite [Enzner, 2006, Gustafsson et al., 2002, Habets et al., 2008, Martin and Altenhoner, 1995] where a unique postfilter is used to tackle noise and echo together. Combined approaches to AEC and echo postfilter present the advantage of reducing both the computational complexity and the signal delay. The objective of the work reported in this thesis is to improve this architecture and the interactions between the different modules so as to propose an more efficient speech enhancement scheme.

2.3 Assessment tools

Good speech quality in phone conversation is possible thanks to the numerous speech enhancement algorithms implemented in devices. Prior to their implementation within phones, we need to evaluate and compare the differents algorithms. This assessement is usually done in two steps: first through objective metrics and later on through subjective tests. Objective metrics report the mathematical performance of the algorithm under test. Subjective tests report the perceived speech quality.

A variety of assessment tools for speech enhancement can be found in the litterature. Only a selection of tools used in this thesis are presented here.

2.3.1 Objective metrics

Echo cancellation algorithms aim to suppress the echo signal recorded by the microphone while NR aims to suppress the noise. It therefore makes sense to assess these algorithms in terms of the amount of perturbation (noise or echo) suppressed. Echo suppression is assessed in terms of echo return loss enhancement (ERLE) whereas NR is assessed in terms of noise attenuation (NA). Perturbation attenuation is measured over adjacent windows of N samples:

$$D(m) = 10.\log_{10}\Big(\frac{\sum_{l=1}^{N} s_{in}^2(mN+l)}{\sum_{l=1}^{N} s_{out}^2(mN+l)}\Big) \tag{2.26}$$

where $D(m)$ stands for ERLE or NA in dB, N is the block size, s_{in} and s_{out} are the unprocessed and processed signals respectively. ERLE and NA are both computed according to Equation 2.26. However, ERLE is measured during echo only periods while NA is measured during noise only periods. Both should be positive and as high as possible.

The echo postfiltering and NR are both achieved in the frequency or subband domain and can sometimes distort the useful signal. Speech distortion can be assessed in terms of cepstral distance (CD) and speech attenuation (SA). The CD can be measured as in [Fingscheidt and Suhahi, 2007] i.e. between the clean speech $s(n)$ and the weighted speech signal $\bar{s}(n)$ as follows:

$$\begin{aligned} C_s(n) &= IDFT\left\{\log|DFT(\mathbf{s(n)})|\right\} \\ CD(m) &= \sqrt{\textstyle\sum_{l=1}^{N}[C_s(m) - C_{\bar{s}}(m)]^2}. \end{aligned} \tag{2.27}$$

The weighted speech signals $\bar{s}(n)$ is obtained with a method similar to [Fingscheidt and Suhahi, 2007]. When processing degraded speech signals, the updated spectral gains $W(k,i)$ are stored. These gains are applied to the clean near-end speech $s(n)$ to obtain the weighted speech signal $\bar{s}(n)$. Similarly the SA can be measured as the attenuation between the clean speech $s(n)$ and the weighted speech signal $\bar{s}(n)$ [Fingscheidt and Suhahi, 2007] as follows:

$$SA(m) = 10\log_{10}\frac{\sum_{l=1}^{N} s^2(mM+l)}{\sum_{l=1}^{N} \bar{s}^2(mN+l)} \quad \text{in dB.} \tag{2.28}$$

Using the weighted speech signal $\bar{s}$ instead of $\hat{s}$ in the computation of CD and SA permits to focus on the distortion of the useful signal by discarding the distortion that might be to residual perturbation. The CD gives an information about the distortions (i.e. musical noise) of the processed signal whereas the SA reflects the attenuation of the useful signal. Ideally, the CD and SA should be equal to 0 and 0dB respectively.

The echo postfiltering and NR gains both require the estimation of the perturbation PSD. Good echo or noise suppression also depends on the accuracy of the estimated PSD. We assess the accuracy of a PSD estimator by the mean of symmetric segmental logarithmic error [Jeub et al., 2012] which can be expressed as follows:

$$logErr = \frac{1}{KM}\sum_{k=0}^{K-1}\sum_{i=0}^{M-1}\left|10\log_{10}\left[\frac{\Phi^{zz}(k,i)}{\hat{\Phi}^{zz}(k,i)}\right]\right| \tag{2.29}$$

where K is the number of frames and M is the number of subbands or frequency bins and z is either the noise or the residual echo. The $logErr$ should be as close as possible to 0.

2.3.2 Subjective tests

Subjective tests come as final validation step for speech enhancement algorithms. They aim at evaluating the perceived speech quality. Most popular subjective test methodologies are defined in the ITU-T Recommendations [ITU-T, 1996b]. We distinguish two families of tests: comparative tests and absolute tests. In comparative tests, the tester is presented two occurrences of the same speech signals that have been processed differently. The tester has to rate both samples. For absolute tests, the tester is presented one sample that he has to evaluate. Subjective tests are very costly as they required a large number of participants and time.

In this thesis, we refer to informal listening tests which are performed by one or two audio experts. Although they are less significant than subjective tests, they still report the subjective quality of the signals.

2.4 Conclusions

In this chapter, we presented the problem due the acoustic echo and noise in a phone conversation. We also explain that the annoyance related to echo and noise increases as the the level of the disturbance increases which justifies the necessity of speech enhancement algorithms.

The main focus of this thesis being the echo problem, a broad presentation of state-of-the-art echo control algorithms has been presented. Most echo control systems are based on adaptive filtering followed by residual echo suppression. Echo suppression performance can be improved by using control such as DTD or synchronization modules. Part of this chapter is dedicated to NR algorithms. NR mostly consists of an attenuation gain applied to the noisy signal, the key elements of the NR being the gain rule and the noise estimate.

The last part of this chapter deals will assessment tools that are commonly used to evaluate speech enhancement algorithms.

The noise reduction and echo postfiltering both operate in the subband or frequency domain. Computational complexity reduction can be obtained by combining both modules. In the next chapter, we study the interest of different frequency and subband domain filtering methods for combined noise reduction and echo postfiltering.

Part I

Single microphone echo processing

Chapter 3

Frequency and subband domains related filtering methods

Contents

In the previous chapter we presented problems that arise in sound recording with mobile devices and the necessity to process the recorded signals before their transmission through the communications network. We also presented state-of-the-art speech enhancement algorithms that can be implemented in mobiles devices in order to guarantee satisfactory speech quality.

Noise reduction and echo postfiltering both consist of time-varying attenuation gains that are updated in the frequency or subband domain with the perturbation (noise or residual echo) suppression being applied in the same frequency or subband domain. Filtering in the frequency or subband domain is advantageous because of its computational complexity. Nevertheless, subband filtering suffers from important processing delay [Löllmann and Vary, 2007]. Delay reduction methods consist in filtering the perturbations in the time domain instead of the subband domain. In this case, subband attenuation gains are used to define a time domain filter [Löllmann and Vary, 2007, Steinert et al., 2008].

Frequency domain filtering suffers from time domain aliasing due to circular convolution [Oppenheim and Schafer, 1999]. Time domain aliasing due to circular convolution can be avoided through overlapping frames and/or zero-padding with the aim of approaching linear convolution [Vaidyanathan, 1993]. Linear convolution nevertheless has an important computational complexity. We show how computational complexity of linear convolution can be reduced by introducing a scalable approach to its implementation.

In this chapter, we present filtering methods that can be used in replacement of subband and frequency domain filtering and assess their performance for noise reduction and echo postfiltering. This chapter is organized as follows. In the next section, we describe how conversion from time to subband or frequency domain is achieved. Section 3.2 addresses problems due to filtering in the subband domain and alternative time domain filtering methods that can be used. In Section 3.3, we present the problems of frequency domain filtering and propose some methods to overcome these problems. Section 3.4 presents a comparative assessment of the different filtering methods for noise reduction and echo postfiltering.

The hereby contributions have been published in conferences proceedings [**Yemdji** et al., 2010a,b, 2011].

3.1 Short time Fourier transform

The short-time Fourier transform (STFT) of a discrete signal $e(n)$ is defined as follows [Allen and Rabiner, 1977]:

$$e(k,i) = \sum_{r=-\infty}^{+\infty} e(kR-r)\cdot p(r)\cdot \exp\Big(-\frac{2\pi}{M}ji(kR-r)\Big) \tag{3.1}$$

where k is the frame index, i is the frequency index, j is the imaginary unit, M is the number of frequency bins, R is the blocksize (i.e. $e(n)$ is processed by block of R samples) and p is the analysis window. Assuming the analysis window is of finite length L, $e(k,i)$ becomes:

$$e(k,i) = \sum_{r=0}^{L-1} e(kR-r)\cdot p(r)\cdot exp\Big(-\frac{2\pi}{M}ji(kR-r)\Big). \tag{3.2}$$

Equation 3.2 can be interpreted in two different ways. These interpretations have a direct effect on the STFT implementations and on the window constraints.

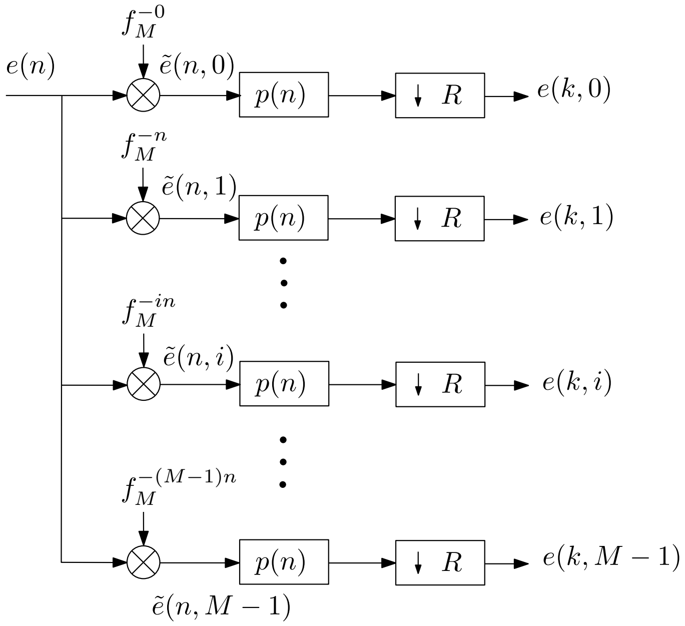


Figure 3.1: DFT-modulated filter bank structure

3.1.1 From STFT to filter bank structure

By rewriting Equation 3.2 in the form of a convolution as follows,

$$e(k,i) = \sum_{r=0}^{L-1} \tilde{e}(kR - r) \cdot p(r) \tag{3.3}$$

$$e(k,i) = (\tilde{e} \star p)(kR) \quad \text{where} \quad \tilde{e}(n,i) = e(n) \cdot f_M^{-in} \quad \text{and} \quad f_M^{-in} = exp\Big(-\frac{2\pi}{M}jin\Big) \tag{3.4}$$

the STFT can be interpreted as an analysis filter-bank. Equation 3.3 shows that $e(k,i)$ is obtained through a convolution as illustrated in Figure 3.1. The input signal $e(n)$ is modulated by a complex exponential. The modulation by the complex exponential corresponds to a frequency shift of the input signal spectrum towards the central frequency $i = 0$. The modulated signal $\tilde{e}(n,i)$ is then lowpass filtered by $p(n)$ to obtain the subband signal $e(k,i)$. The structure illustrated in Figure 3.1 is referred to as the discrete Fourier transform (DFT)-modulated analysis filter-bank [Vaidyanathan, 1993]. In the filter-bank interpretation, R plays the role of the downsampling parameter. The structure illustrated in Figure 3.1 is referred to as the discrete Fourier transform (DFT)-modulated analysis filter-bank and the window $p(n)$ is called prototype filter [Crochiere and Rabiner, 1983, Vaidyanathan, 1993].

In a similar manner, the inverse STFT can be seen as a synthesis filter bank. To synthesize the full band signal, the first step consists in convolving each subband signal with the synthesis filter (or window). The M convoluted signals are then modulated by the appropriate complex exponential before being summed to form the fullband signal.

According to Figure 3.1, the M sub bands signals are obtained by M convolutions. The structure in Figure 3.1 could be more efficiently implemented through PolyPhase Network (PPN) implementation. The derivation of the PPN implementation can be found in [Vaidyanathan, 1993]. With the PPN implementation, subbands signals are obtained

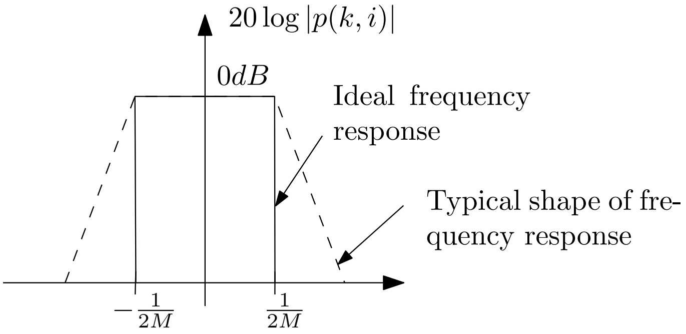


Figure 3.2: Filter bank structure

through one convolution and one Fourier transform. Later on in our implementation, we will the PPN structure.

The resulting subband signal $e(k,i)$ has a lower frequency bandwidth of ${}^{2\pi}/_{M}$. To avoid frequency domain aliasing due to downsampling, the subsampling factor R is constrained to be lower or equal to the number of subbands M [Crochiere and Rabiner, 1983]. The only constraint on the downsampling factor value is not a sufficient condition to avoid aliasing: the subband filters has to be designed such that there is no overlap between two adjacent subband signals. In the case of the DFT-modulated filter bank, adjacent subband signals $e(k,i)$ and $e(k,i+1)$ are obtained by respectively low-pass filtering the modulated signal $\tilde{e}(n,i)$ and $\tilde{e}(n,i+1)$ with the prototype filter $p(n)$. The adjacent subband signals $e(k,i)$ and $e(k,i+1)$ do not overlap if the frequency response of the prototype filer is an ideal lowpass filter (i.e. rectangular window in the frequency domain) which we illustrate in Figure 3.2. In practice it is impossible to design such a lowpass filter. In consequence, when reconstructing the fullband signal through with the synthesis filter, some aliasing will be introduced in the output signal.

3.1.2 From STFT to overlap add

The STFT as defined in Equation 3.2 can also be interpreted as a DFT comprising overlap-add (OLA). In this case, $e(k,i)$ is seen as the DFT of a windowed version of the signal $e(n)$:

$$\begin{aligned} e(k,i) &= \sum_{r=0}^{L-1} \Big[e(kR-r) \cdot p(r) \Big] \cdot exp\Big(-\frac{2\pi}{M} ji(kR-r) \Big) & (3.5) \\ &= \sum_{r=0}^{L-1} e_p(kR-r) \cdot exp\Big(-\frac{2\pi}{M} ji(kR-r) \Big) & (3.6) \end{aligned}$$

where $e_p(n)$ denotes the windowed version of $e(n)$. The input signal is processed by block of R new samples to which $(L-R)$ samples from the previous frame(s) are appended to obtain a frame of L samples. The frame of L is then windowed by multiplication with p to obtain $e_p(n)$. Finally, $e(k,i)$ is obtained as the DFT of the signal $e_p(n)$. So for the OLA method, the STFT is seen as the DFT of successive windowed frames.

In contrast to the filter-bank view, the perfect reconstruction in the OLA view leads to a constraint on the window $p(n)$ in the time domain [Crochiere and Rabiner, 1983]. The sum of time shifted version of $p(n)$ should be equal to 1 as illustrated in Figure 3.3.

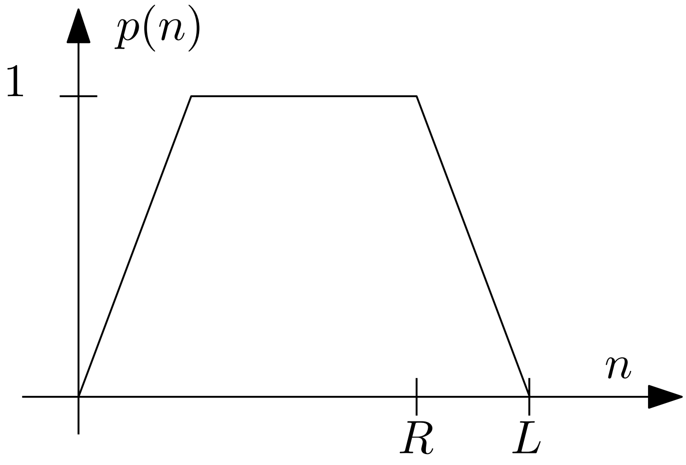


Figure 3.3: Illustration of the overlap add constraint

In the remaining of this chapter, we differenciate the subband and frequency analysis by denoting p_{fb} the prototype filter of subband processing and p_{ola} the overlapping window for the frequency domain processing.

3.2 Filter bank related filtering methods

Speech enhancement generally use the subband signal to compute an attenuation gain which aims are suppressing perturbations (i.e. noise). The focus of this chapter is the perturbation suppression itself (i.e. filtering). In the following we present filtering methods that can be used to suppress the perturbation from the input signal.

Conversion of input signals from time to subband domain takes place as described in Section 3.1.1. The prototype filter considered expresses as follows:

$$p_{fb}(n) = \frac{1}{M} \cdot \text{sinc}\Big[\frac{2\pi}{M}\Big(n - \frac{L}{2}\Big)\Big] \cdot p_L(n), \tag{3.7}$$

where p_L is a Hamming window also of length L. We denote $W(k,i)$ the attenuation which can be applied to the disturbed signal $e(k,i)$ in the subband domain through multiplication as well as in the time domain through convolution. For time domain filtering, the subband gains $W(k,i)$ are converted into a finite impulse response (FIR) filter. The resulting FIR filter $w(n)$ has the same frequency response as $W(k,i)$.

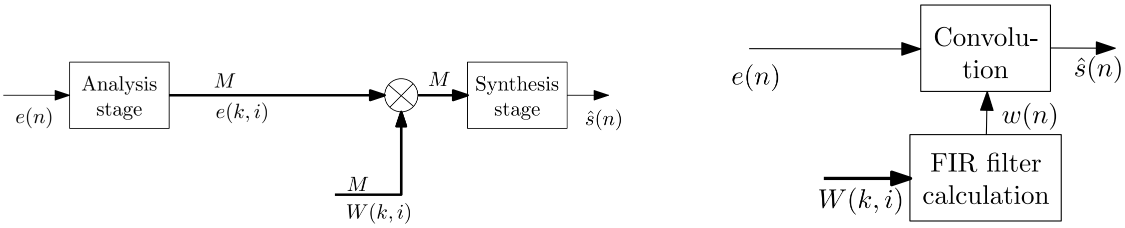
Subband filtering is presented in Section 3.2.1. In Section 3.2.2, we present time domain filtering and approaches that can be used to compute the FIR filter.

3.2.1 Subband domain weighting

Subband filtering consists in applying the i^{th} subband gain $W(k,i)$ to the i^{th} subband signal $e(k,i)$ as a multiplicative factor (see Figure 3.4(a)):

$$\hat{s}(k,i) = W(k,i) \cdot e(k,i). \tag{3.8}$$

The M subband signals $\hat{s}(k,i)$ are processed through a synthesis filter bank to recover the fullband estimate of the useful signal $\hat{s}(n)$. In the following, we refer to this method as the ASFB method which stands for analysis synthesis filter-bank method.



(a) sub-band filtering (b) Time domain filtering

Figure 3.4: Filtering methods. Bold lines represents subband domains

The ASFB method is interesting for its simplicity: filtering simply consists of a multiplication. Nevertheless, this method suffers from frequency domain aliasing and significant signal delay. The overall ASFB system introduces a signal delay of $L-1$ samples, where L is the length of the prototype filter. As explained in Section 3.1.1, subband domain aliasing can be reduced with the use of an appropriate prototype filter. Typically, long prototype filters permit to achieve good attenuation of the aliasing components and but lead to significant signal delay [Löllmann, 2011]. Moreover, the longer the prototype filter is, the more the computational complexity of the filter-bank is high.

As explained in Section 2.2.3, signal delay and computational complexity are important constraints in speech enhancement for mobile devices. Computational complexity and delay reduction can be reduced achieve using short prototype filters. Further delay reduction can be achieved by using time domain filters instead of the whole ASFB structure. In the following section, we discuss approaches to filter the disturbed signal in the time domain.

3.2.2 Time domain filtering

Figure 3.4(b) shows an equivalent filtering scheme when a time-domain filter is used for perturbation suppression. The time varying attenuation gains $W(k,i)$ are used to determine an FIR filter. Besides the signal delay reduction, time domain filtering also permits to avoid problems due to frequency domain aliasing. In contrast to the ASFB method, no synthesis filter-bank is required. The filtering is applied through convolution with the aliasing-free input signal $e(n)$.

The FIR filter can be calculated according to 3 different conversion methods. To avoid phase distortions and to ensure a constant signal delay, we ensure that these filters are linear phase filters [Crochiere and Rabiner, 1983]. We now present the 3 conversion methods considered.

3.2.2.1 Filter bank equalizer (FBE)

The FBE was introduced in [Löllmann and Vary, 2007] and is the mathematical time domain equivalent of the analysis filter-bank with synthesis through summation. The FBE is expressed as follows:

$$w(n) = p_{fb}(n) \cdot \tilde{w}(n) \quad \text{with} \quad \tilde{w}(n) = \frac{1}{M}\sum_{i=0}^{M-1} W(k,i) \cdot exp\Big(\frac{2\pi}{M}jik\Big) \tag{3.9}$$

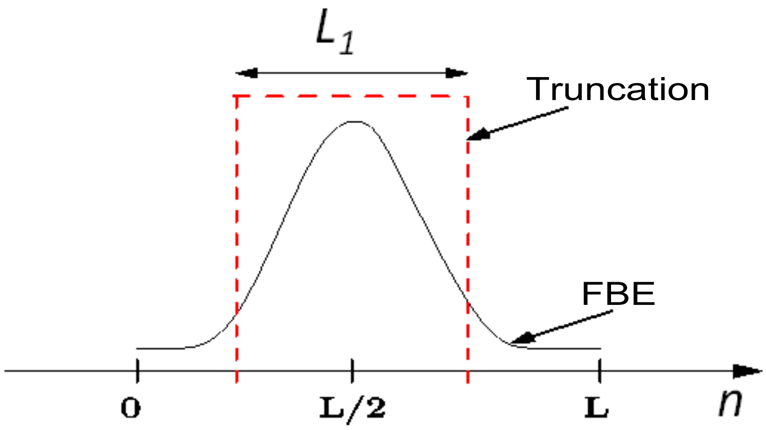


Figure 3.5: Illustration of the determination of the LDF

where $p_{fb}(n)$ is the prototype filter of the sub-band analysis stage and $\tilde{w}(n)$ is the IDFT of the spectral gains $W(k,i)$. The FBE process introduces a signal delay of $(L-1)/2$ samples (with L being the length of the $p(n)$), that is half the delay introduced by the ASFB method.

3.2.2.2 Low delay filter (LDF)

Although the FBE has lower signal delay than the corresponding ASFB, smaller signal delays can be achieved by approximating the FBE by a lower degree filter [Löllmann and Vary, 2007, 2009]. The LDF is obtained by truncating the FBE with a window of length L_1 with $L_1 < L$ as illustrated in Figure 3.5. The window used can be an arbitrary window of L_1-taps. The window and the truncation should be chosen to maintain linear phase properties. For this, as the FBE has linear phase, the window used for the LDF should be symmetric about $L/2$.

3.2.2.3 Inverse Discrete Fourier transform (IDFT)

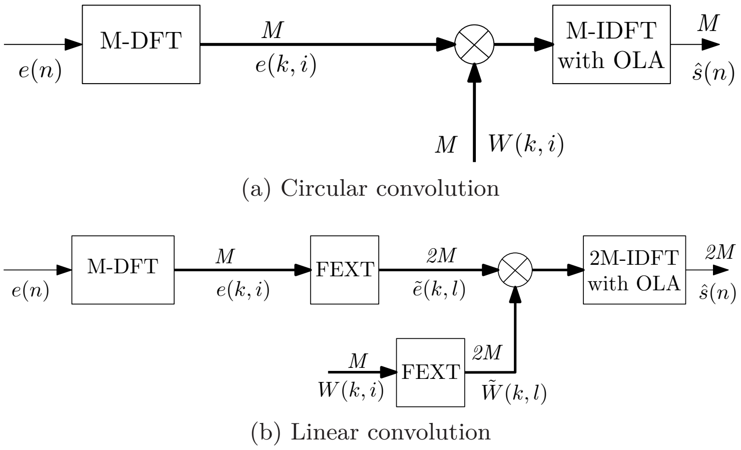
A more intuitive approach consists in defining the FIR filter simply as the IDFT of the updated subband gain factors [Hänsler and Schmidt, 2000]. The IDFT of a positive sequence corresponds to an even symmetric sequence in the time domain. The IDFT of the subband gains $W(k,i)$ corresponds to a non-causal zero phase filter. A causal filter is obtained by applying a temporal shift of $(M-1)/2$

$$w(n) = \tilde{w}(n - \frac{M-1}{2}) \tag{3.10}$$

The temporal shift corresponds to a linear modification of the phase in the frequency domain. The causal filter has linear phase and will thus introduce a group delay of (M-1)/2. The linear phase property is important in speech processing properties in order to avoid phase distortions.

3.3 Discrete Fourier transform related filtering methods

In some cases, signals are processed in the frequency domain. Conversion of input signals from time to frequency domain takes place as described in Section 3.1.2. $W(k,i)$



(a) Circular convolution

(b) Linear convolution

Figure 3.6: Block processing in the frequency domain. Bold lines represent the frequency domain.

still denotes the time varying attenuation gains and is used to process $e(k,i)$ through multiplication.

Filtering in the frequency is advantageous for its computational simplicity but is subject to aliasing resulting from circular convolution. Linear convolution can be used to avoid this aliasing problem but is nevertheless very computationally demanding. In the following, we present present the problem due to circular convolution and approaches to implement the linear convolution. A novel implementation approach for linear convolution in the frequency domain is also introduced.

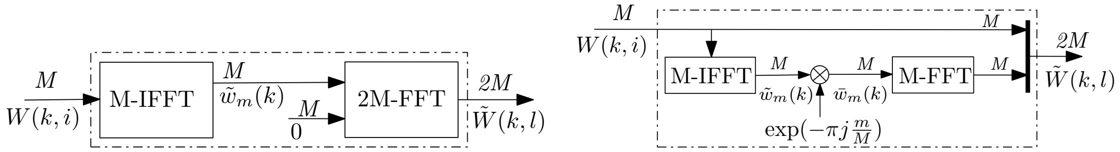
3.3.1 Circular convolution

Figure 3.6(a) illustrates circular convolution which is the method used for filtering signals in the frequency domain. The filtered signal is obtained through a bin-by-bin multiplication of the M frequency domain components $e(k,i)$ with the gains $W(k,i)$. As shown in Figure 3.6(a), filtering an M-point signal with an M-tap filter produces an M-point signal instead of an $(2M-1)$-point signal as would normally result from convolution in the time domain. This observation implies that the filtering of signals in the frequency domain as shown in Figure 3.6(a) introduces distortion in processed signals [Oppenheim and Schafer, 1999]. Distortions introduced by circular convolution result from time domain aliasing. Details regarding the cause of this aliasing can be found in 3.A.

3.3.2 Linear convolution

Figure 3.6(b) depicts the scheme of linear convolution in the frequency domain. Here, the frequency domain signals $e(k,i)$ and gains $W(k,i)$ are processed through a frequency resolution extension (FEXT) block prior to the filtering operation. Zero-padding takes place withtin the FEXT block and is used to increase the number of frequency bins. As illustrated in Figure 3.7(a), the FEXT block operates in two steps:

- The M frequency bin input signals $W(k,i)$ are converted into the time domain with an IFFT of length M (M-IFFT) to obtain $\tilde{w}_m(k)$ where m is the tap index of the impulse response and ranges from 0 to $M-1$.



(a) Frequency extension definition

(b) Efficient FEXT implementation

Figure 3.7: FEXT implementations

- The time domain signal $\tilde{w}_m(k)$ is zero-padded with at least M zeros to obtain a signal of length $2M$ and reconverted into the frequency domain through a $2M$-FFT to obtain $\tilde{W}(k,l)$ with l being the "new" frequency bin index ranging from 0 to $2M-1$.

Returning to Figure 3.6(b), we now have $2M$ frequency bins instead of M. As a results, the filtered signal has $2M$ samples.

The merit of linear convolution over circular convolution is that it does not introduce distortion since it is equivalent to filtering in the time domain. Linear convolution, as described here, requires additional DFTs of size $2M$. Its major disadvantage is its increased computational load and memory requirement which is due to the use of the FEXT module. In most real time systems, the computation of FFTs is based on exponential function that are defined for a given FFT size and stored in the memory. In addition, the filtered signal is longer: it has $2M$ points instead of M points for the linear convolution. Proper reconstruction of the fullband signal implies that this vector be stored in the memory of the system. The additional FFTs required by the linear convolution mean an increase of the required memory. In the following, we focus on efficient methods to implement the FEXT module.

3.3.3 Link between circular and linear convolution

The approach presented in this section was introduced in [Marin-Hurtado and Anderson, 2010]. According to Figure 3.7(a), the frequency domain gains $\tilde{W}$ used in the linear convolution are defined as follows:

$$\tilde{W}(k,l) = \sum_{m=0}^{2M-1} \tilde{w}_m(k) \cdot \exp\left(-2\pi j \frac{ml}{2M}\right) = \sum_{m=0}^{M-1} \tilde{w}_m(k) \cdot \exp\left(-2\pi j \frac{ml}{2M}\right). \tag{3.11}$$

The summation terms between $m = M$ and $m = 2M-1$ are omitted since $\tilde{w}_m(k) = 0$ for this interval. By splitting Equation 3.11 into two, for even and odd values of l, we obtain:

$$\tilde{W}(k,l=2i) = \sum_{m=0}^{M-1} \tilde{w}_m(k) \cdot \exp\left(-2\pi j \frac{mi}{M}\right) = W(k,l/2) \tag{3.12}$$

$$\tilde{W}(k,l=2i+1) = \sum_{m=0}^{M-1} \tilde{w}_m(k) \cdot \exp\left(-2\pi j \frac{m(i+1/2)}{M}\right), \tag{3.13}$$

where i is an integer ranging from 0 to $M-1$. Equation 3.12 shows that for even values of l, $\tilde{W}(k,l)$ is equal to $W(k,l/2)$. For odd values of l (Equation 3.13), $\tilde{W}(k,l)$ is a discrete Fourier transform except for the unusual exponential term. Equation 3.13 can be rewritten as

$$\tilde{W}(k, l = 2i+1) = \sum_{m=0}^{M-1} \bar{w}_m(k) \cdot \exp\left(-2\pi j \frac{mi}{M}\right) \tag{3.14}$$
$$\text{where} \quad \bar{w}_m(k) = \tilde{w}_m \cdot \exp\left(-\pi j \frac{m}{M}\right).$$

From Equation 3.14, it is apparent that $\tilde{W}(k, l = 2i+1)$ can be computed through an FFT algorithm as shown in Figure 3.7(b). The FEXT implementation then requires two M-point FFTs and M complex multiplications. By considering Figure 3.7(b), we see that the $2M$-points FFT is discarded from the FEXT implementation. The computational complexity of the FFT increases with its size. This implementation approach to linear convolution has a reduced computational complexity compared to original linear convolution.

3.3.4 Alternative interpretation of the linear convolution

In this section we introduce a new implementation of the linear convolution. Our approach to implement the FEXT exploits the fact that zero-padding in the time domain is equivalent to interpolation in the frequency domain and vice-versa. In the following we introduce the relationship between $\tilde{W}(k, l)$ and $W(k, i)$ using the FFT and IFFT definitions. As the impulse response $\tilde{w}_m(k)$ is the IFFT of $W(k, i)$ Equation 3.11 can be rewritten as;

$$\tilde{W}(k, l) = \sum_{m=0}^{M-1} \Big[\left[\frac{1}{M} \sum_{i=0}^{M-1} W(k, i) \exp\left(2\pi j \frac{mi}{M}\right) \right] \exp\left(-2\pi j \frac{ml}{2M}\right) \Big] \tag{3.15}$$

$$= \frac{1}{M} \sum_{i=0}^{M-1} W(k, i) \sum_{m=0}^{M-1} \left(\exp\left(\frac{2\pi}{2M} j\,(2i - l)\right) \right)^m. \tag{3.16}$$

The sum of exponential in Equation 3.16 is equal to:

$$\sum_{m=0}^{M-1} \left(\exp\left(\frac{2\pi}{2M} j\,(2i - l)\right) \right)^m = \begin{cases} M & \text{if } l \text{ is even and } l = 2i \\ 0 & \text{if } l \text{ is even and } l \neq 2i \\ \frac{1 - \exp\left(\frac{2\pi}{2M} j M (2i-l)\right)}{1 - \exp\left(\frac{2\pi}{2M} j (2i-l)\right)} = \frac{2}{1 - \exp\left(\frac{2\pi}{2M} j (2i-l)\right)} & \text{if } l \text{ is odd.} \end{cases} \tag{3.17}$$

Inserting Equation 3.17 into Equation 3.16 leads to:

$$\tilde{W}(k, l = 2i) = W(k, i) \tag{3.18}$$

$$\tilde{W}(k, l = 2i+1) = \frac{1}{M} \sum_{i=0}^{M-1} \frac{2}{1 - \exp\left(\frac{2\pi}{2M} j\,(2i - l)\right)} W(k, i). \tag{3.19}$$

Equation 3.18 confirms Equation 3.12. We denote $z_{l,i}$ the weighting factor of $W(k, i)$ in Equation 3.19:

$$z_{l,i} = \frac{1}{M} \frac{2}{1 - \exp\left(\frac{2\pi}{2M} j\,(2i - l)\right)}. \tag{3.20}$$

One can easily verify that $z_{l,i}$ is such that $z_{l+2,i} = z_{l,i-1}$

$$z_{l+2,i} = \frac{1}{M} \frac{2}{1 - \exp\left(\frac{2\pi}{2M} j\,(2i - (l+2))\right)} \tag{3.21}$$

$$= \frac{1}{M} \frac{2}{1 - \exp\left(\frac{2\pi}{2M} j\,(2(i-1) - l)\right)} \tag{3.22}$$

$$= z_{l,i-1}. \tag{3.23}$$

This property of $z_{l,i}$ is of particular interest if we write Equation 3.19 in matrix form:

$$\begin{bmatrix} \tilde{W}_1 \\ \tilde{W}_3 \\ \vdots \\ \tilde{W}_{2M-1} \end{bmatrix} = \begin{bmatrix} z_{1,0} & z_{1,1} & \cdots & z_{1,M-1} \\ z_{3,0} & z_{3,1} & \cdots & z_{3,M-1} \\ \vdots & & & \\ z_{2M-1,0} & z_{2M-1,1} & \cdots & z_{2M-1,M-1} \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{M-1} \end{bmatrix}. \tag{3.24}$$

Using the properties of $z_{l,i}$ mentioned above, Equation 3.24 then becomes:

$$\begin{bmatrix} \tilde{W}_1 \\ \tilde{W}_3 \\ \vdots \\ \tilde{W}_{2M-1} \end{bmatrix} = \begin{bmatrix} z_{1,0} & z_{1,1} & \cdots & z_{1,M-1} \\ z_{1,M-1} & z_{1,0} & \cdots & z_{1,M-2} \\ \vdots & & & \\ z_{1,1} & z_{1,2} & \cdots & z_{1,0} \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{M-1} \end{bmatrix}. \tag{3.25}$$

Equation 3.25 shows that the matrix Z formed from the weighting factors $z_{l,i}$, is circulant. A well known property of circulant matrices is that they are diagonalizable by Fourier matrices [Petersen and Pedersen, 2012]:

$$Z = FDF^{-1}, \tag{3.26}$$

where D is a diagonal matrix such that $D = \operatorname{diag}(F\mathbf{z})$, $\mathbf{z}$ is the vector formed by the elements of the first column of Z and F is a Fourier matrix (i.e. composed of exponential terms $f_M^{ik} = e^{-2\pi jik/M}$). Equation 3.25 can then be rewritten as:

$$\tilde{\mathbf{W}} = ZW = FDF^{-1}\mathbf{W}. \tag{3.27}$$

This approach to linear convolution can also be computed through the scheme illustrated in Figure 3.7(b). In Equation 3.27, the product $F^{-1}\mathbf{W}$ corresponds to the IFFT of the spectral gains $W(k,i)$ and is equal to $\tilde{w}_m(k)$. Experiments verified that the diagonal matrix D is indeed composed of the same terms as the exponent term illustrated in Figure 3.7(b). Meaning that the product $DF^{-1}\mathbf{W}$ leads to $\bar{w}_m(k)$ since $F^{-1}\mathbf{W} = \tilde{w}_m(k)$. Lastly, the product by the Fourier matrix achieves the Fourier transform of $\bar{w}_m(k) = DF^{-1}\mathbf{W}$.

3.3.4.1 Frequency resolution extension with reduced computational complexity

Our approach (Section 3.3.4) to linear convolution corroborates the work in [Marin-Hurtado and Anderson, 2010] but has the distinct advantage of scalability for managing computational complexity which we describe here. In this section we consider the case where $\tilde{W}(k,i)$ is computed according to Equation 3.19, i.e. through multiplication of $W(k,i)$ by the weighting function $z_{l,i}$.

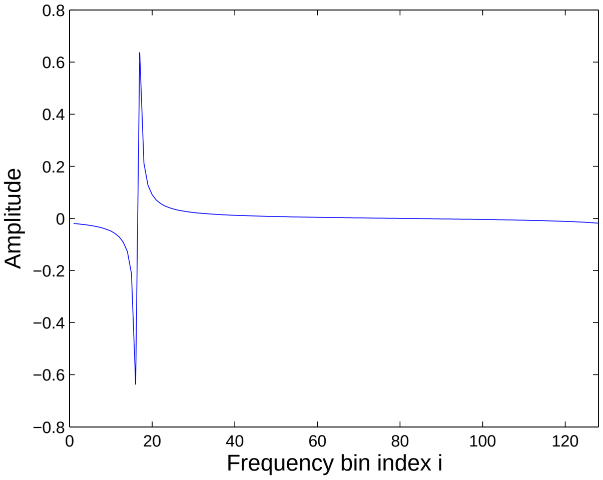


Figure 3.8: Imaginary part of the weighting function $z_{l,i}$ for $l = 33$ (i.e. normalized frequency $l/2M$) and $M = 128$

The imaginary part of the weighting function $z_{l,i}$ for $l = 33$ is depicted in Figure 3.8. The real part of $z_{l,i}$ is not shown because it is constant for all values of i: $\mathrm{Re}(z_{l,i}) = 1/M$. The plot in Figure 3.8 shows that $z_{l,i}$ does not equally weight the spectral gains $W(k,i)$ in the computation of $\tilde{W}(k,l)$. More specifically, the closer the normalized frequency $\frac{i}{M}$ is to the normalized frequency $\frac{l}{2M}$, the more $W(k,i)$ influences the value of $\tilde{W}(k,l)$ (and vice-versa). Although, $z_{l,i}$ is complex and does not uniformly weight $W(k,i)$, it is of interest to note that a given value of l, $z_{l,i}$ constitutes a normalized weigthing function:

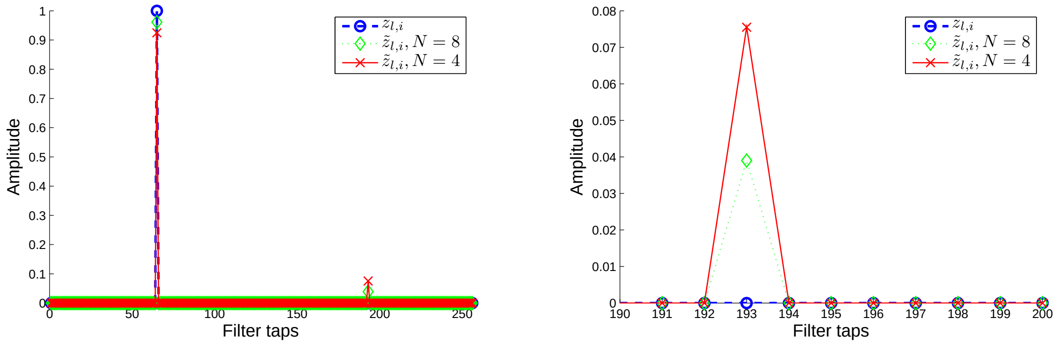
$$\sum_{i=0}^{M-1} z_{l,i} = 1. \tag{3.28}$$

To reduce the computational complexity related to the computation of $\tilde{W}(k,l)$, one can use a truncated version of $z(k,i)$ as a weighting function. We denote $\tilde{z}_{l,i}$ the truncated version of $z_{l,i}$. The weighting function $\tilde{z}_{l,i}$ is truncated so as to include only N points (with $N < M$) centered on the peak of $z_{l,i}$. We denote $\tilde{\tilde{W}}(k,l)$ the attenuation gain obtained with the truncated interpolation function $\tilde{z}_{l,i}$. $\tilde{\tilde{W}}(k,l)$ is computed similarly as in Equation 3.19:

$$\tilde{\tilde{W}}(k, l = 2i+1) = \sum_{i=0}^{M-1} \tilde{z}_{l,i} \cdot W(k,i) = \sum_{i}^{N} \tilde{z}_{l,i} \cdot W(k,i). \tag{3.29}$$

where only N terms are considered in the summation since $\tilde{z}_{l,i}$ has $M - N$ terms equal to 0. The computation of $\tilde{\tilde{W}}(k,l)$, as in Equation 3.29 requires less operations. For all the spectral gains $\tilde{\tilde{W}}(k,l)$, this computation requires $N \cdot M$ multiplications and $(N-1) \cdot M$ summations.

The scalability comes from the fact that one can choose the value of N according to the computational load of the system. The bigger N, the closer the resulting $\tilde{\tilde{W}}(k,l)$ will be to $\tilde{W}(k,l)$. Compared to the optimum computation of Equation 3.27, the use of $\tilde{z}_{l,i}$ is computationally advantageous when $N \leq 2 \cdot log_2(M)$ if we assume that an M-point FFT



(a) Impulse response

(b) Zoom on the impulse response

Figure 3.9: Impulse response for FEXT different weighting function configurations

has a computational complexity of $Mlog_2(M)$ (e.g. for $M = 256$, N should be be lower of equal to 16).

The truncation of the weighting function $z_{l,i}$ to reduce the computational complexity of the system may introduce some distortion since the resulting approximation $\tilde{z}_{l,i}$ contains less information than the original function. In order to evaluate the impact of such approximation, in the following we analyze the impact of the truncation on spectral gains and analysis-synthesis of speech signals. Thorough experiments in a real perturbation suppression scheme are presented later on in Section 3.4.

3.3.4.2 Impact of FEXT optimization on filters

We evaluate the impact of the truncation by comparing the impulse response and the frequency response of a filter obtained with the full weighting function to that obtained with a truncated weighting function. In the test reported here, spectral gains $W(k,i)$ are all set to 0dB and M is set to 128. These spectral gains are processed by the FEXT to obtain new spectral gains. Figure 3.9(a) shows the impulse responses obtained when FEXT uses the full weighting function $z_{l,i}$ or two truncated weighting functions $\tilde{z}_{l,i}$ (of length $N = 8$ and $N = 4$ respectively). Values of N are chosen such that the resulting system has lower computational complexity than the linear convolution (i.e. $N \leq 2 \cdot log_2(M)$).

We observe that impulse responses obtained with the truncated weighting functions do not exactly match the reference impulse response which, in this case, is composed of a unique peak of unit amplitude. In the case of the approximated impulse responses, the peak amplitude is slightly lower than 1 (0.9802 for $N = 8$ and 0.9244 for $N = 4$ respectively). Moreover, as we can see on Figures 3.9(a) and 3.9(b), both approximations have a second peak which is small compared to the main peak but whose amplitude increases with decreasing N (0.04 for $N = 8$ and 0.08 for $N = 4$ respectively). When comparing the spectrum of the approximated impulse responses $\tilde{\tilde{W}}(k,l)$ with that of the original spectral gains $W(k,i)$, we observe that $\tilde{\tilde{W}}(k,l)$ contains a small ripple. The spectral gains are no longer equal to 0dB (as it is the case for the input spectral gains $W(k,i)$ or $\tilde{W}(k,l)$). The spectral gains of the odd frequency bins are constant (i.e. -0.35dB for $N = 8$ and -0.7dB for $N = 4$). The spectral gains of the even frequency bins are, as expected, equal to 0dB as they do not require any computation. The effects observed on the filter can be judged as annoying or negligible depending on the application. With

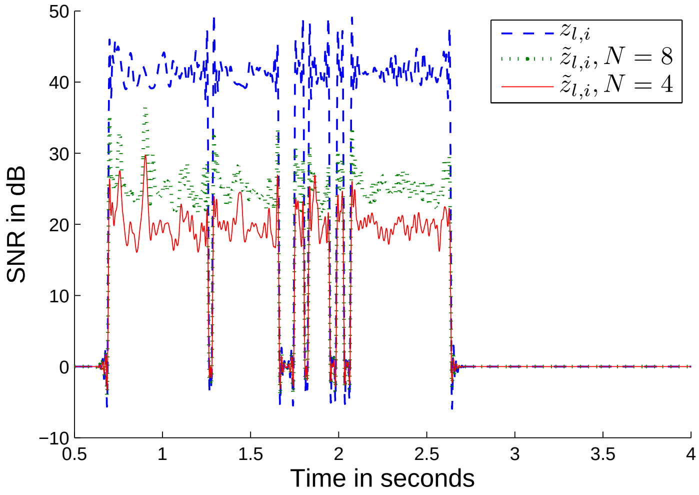


Figure 3.10: Signal-to-noise-ratio (SNR) between input signal and reconstruction error between the input signal and the output of the FEXT module

mobile communications for example, such artifacts may be irrelevant whereas in high quality speech enhancement systems, they may be very annoying.

3.3.4.3 Impact of FEXT optimization on speech signals

Here we report the impact of the optimized FEXT on speech signals. We undertook a similar experiment to that reported in Section 3.3.4.2 but this time using a speech signal as input to the FEXT module. An input speech signal is transformed into the frequency domain through an M-point FFT with overlapping frames of 128 samples (64 new samples and 64 samples from the previous frame) and with Hanning windowing. The number of frequency bins M is set to 128. The obtained spectrum is processed by the FEXT and transformed back into the time domain through a $2M$-point FFT. Except for the FEXT, no processing is performed in the frequency domain. Figure 3.10 shows the signal-to-noise-ratio (SNR) between the input speech and the reconstruction error. We define the reconstruction error as the difference between the input signal and the output signal of the FEXT module. For better analysis, the SNR was forced to zero during speech pauses so that it reflects the impact of the FEXT on the speech signal only. We observed that, without any approximation in the FEXT, the SNR is approximately 40dB during speech periods. The use of a truncated weighting function results in a slight degradation in SNR (about 20dB with $N = 4$). Moreover informal listening tests indicate that these degradations in SNR are not audible. This shows that truncating the weighting function $z_{l,i}$ does not adversely affect speech quality in this case.

3.4 Comparative assessment of the different filtering methods

The subband and frequency domains filtering methods presented in sections 3.2 and 3.3 are assessed for use within a joint noise reduction and echo postfiltering algorithm. Our experimental setup is presented in Section 3.4.1. Results are presented in Section 3.4.2.

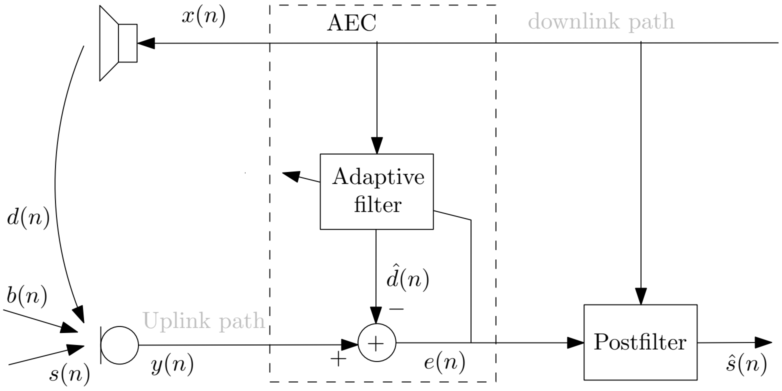


Figure 3.11: Speech enhancement scheme composed of AEC followed by a postfilter. The postfilter aims at reducing both the noise and the residual echo.

3.4.1 Experimental setup

This section is organized as follows. In Section 3.4.1.1, we present the combined noise reduction and echo postfilter used in our investigation. A summary of the filtering methods assessed in presented in Section 3.4.1.2. Lastly Section 3.4.1.3 presents our dataset of test signals.

3.4.1.1 Noise reduction and residual echo suppression

Figure 3.11 shows the echo processing scheme considered in our experiments. The microphone signal $y(n)$ is composed of the near-end speech signal $s(n)$, the echo signal $d(n)$ and the noise signal $b(n)$. The AEC consists of a subband NLMS algorithm with variable stepsize [Degry and Beaugeant, 2008]. The error signal $e(n)$ from the AEC is composed of residual echo $d_r(n)$, the near-end speech $s(n)$ and the noise signal $b(n)$. The postfilter is used to tackle both residual echo and ambient noise. Assuming noise and echo are two additive uncorrelated distance we aimed to suppress, the postfilter spectral gains are defined as the product of the noise reduction W_{noise} and echo suppression gains W_{echo}:

$$W(k,i) = W_{noise}(k,i) \cdot W_{echo}(k,i). \tag{3.30}$$

The noise reduction and echo postfiltering spectral gains are calculated independently. The echo postfilter is updated using a Wiener rule for echo suppression

$$\begin{aligned} W_{echo}(k,i) &= \frac{\xi(k,i)}{1+\xi(k,i)} \\ \text{with} \quad \xi(k,i) &= \beta \frac{\hat{s}^2(k-1,i)}{\hat{\Phi}^{\tilde{d}\tilde{d}}(k-1,i)} + (1-\beta)\cdot \max\big(\frac{e^2(k,i)}{\hat{\Phi}^{\tilde{d}\tilde{d}}(k,i)} - 1, 0\big) \end{aligned} \tag{3.31}$$

where β is the smoothing constant which is set to 0.98 and $\xi(k,i)$ is the (near-end speech) signal-to-(residual) echo ratio (SER) which we estimate through the decision directed approach [Ephraim and Malah, 1983]. The residual echo PSD $\hat{\Phi}^{\tilde{d}\tilde{d}}$ is computed through the cross-correlation method [Beaugeant et al., 1998]:

$$\hat{\Phi}^{\tilde{d}\tilde{d}}(k,i) = \frac{(\Phi^{xe}(k,i))^2}{\Phi^{xx}(k,i)}. \tag{3.32}$$

The noise reduction filter is a low complexity noise reduction algorithm [Degry and Beaugeant, 2008] which is based on the assumption that the amount of noise that should be attenuated is proportional to the signal-to-noise ratio (SNR). The noise reduction gain is expressed as follows:

$$W_{noise}(k,i) = \min(\alpha \cdot \chi^{\lambda}(k,i), 1) \tag{3.33}$$

where α and λ are empirically optimized constants and $\chi(k,i)$ is the SNR which we estimate as:

$$\chi(k,i) = \frac{\Phi^{ee}(k,i)}{\hat{\Phi}^{bb}(k,i)}. \tag{3.34}$$

where $\hat{\Phi}^{bb}$ is the estimate of the noise PSD. The noise PSD is computed through minimum statistics [Martin, 2001].

3.4.1.2 Filtering methods

The updated gains $W(k,i)$ are used to process the degraded speech signal $e(n)$ through the filtering methods described in Sections 3.2 and 3.3. For all filtering methods, the number of subbands or frequency bins M is set to 64.

The four subband related filtering methods of Section 3.2 are considered: ASFB, FBE, LDF and IDFT. For subband related filtering methods, the prototype filter p_{fb} is that specified in Equation 3.7 and its length is set to $L = 128$. The dowsampling factor R is set to 64.

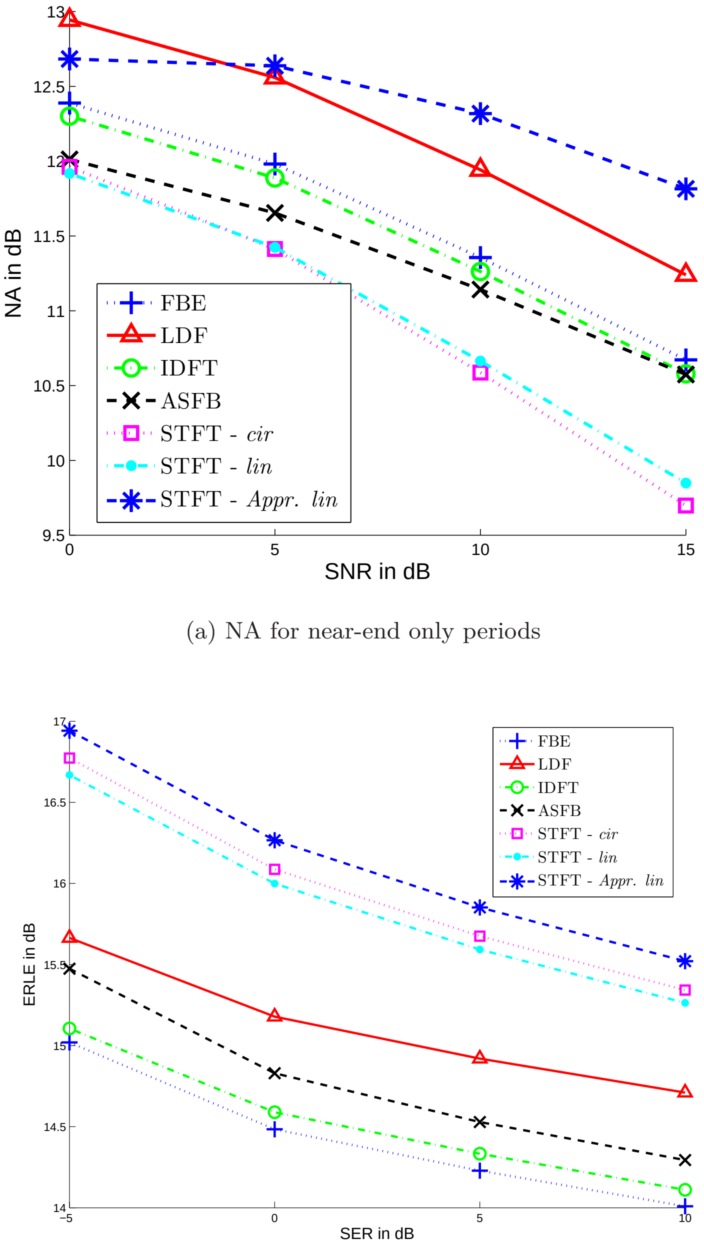
For the filtering methods presented in Section 3.3, conversion from time to frequency domain uses overlapping frames of 128 samples (64 new samples and 64 samples from previous frame). A Hamming window of size $L = 128$ is used as the analysis window. Three filtering methods are considered:

- circular convolution denoted STFT-*cir*
- linear convolution denoted STFT-*lin*
- proposed method with truncated weighting function $\tilde{z}_{l,i}$ and N set to 8 which we denote STFT-*Appr. lin.*

3.4.1.3 Test signals

The microphone signals used in our simulations contain near-end speech only, echo-only and double-talk periods with either car, cafe or babble noise. The echo signal is obtained by convolving the loudspeaker signals with an acoustic path measured from real mobile terminals in an office environment. The loudspeaker and near-end speech levels are both set to -26dB using the ITU-T implementation of the speech voltmeter [ITU-T, 1993] and the different echo and noise levels are also set using the same tool. The SNR ranges from 0 to 15dB while the SER ranges from -5 to 10dB. Our database of degraded speech signals contains 192 sets of microphone and loudspeaker signals of 32s each.

Performance of the different filtering approaches is assessed through objective measurements and informal listening tests. Echo suppression is assessed in terms of echo return loss enhancement (ERLE) and cepstral distance(CD). Noise reduction is assessed in terms of noise attenuation (NA). ERLE, CD and NA are computed as described in Section 2.3.



(a) NA for near-end only periods

(b) ERLE for echo-only periods

Figure 3.12: Perturbation attenuation

3.4.2 Results

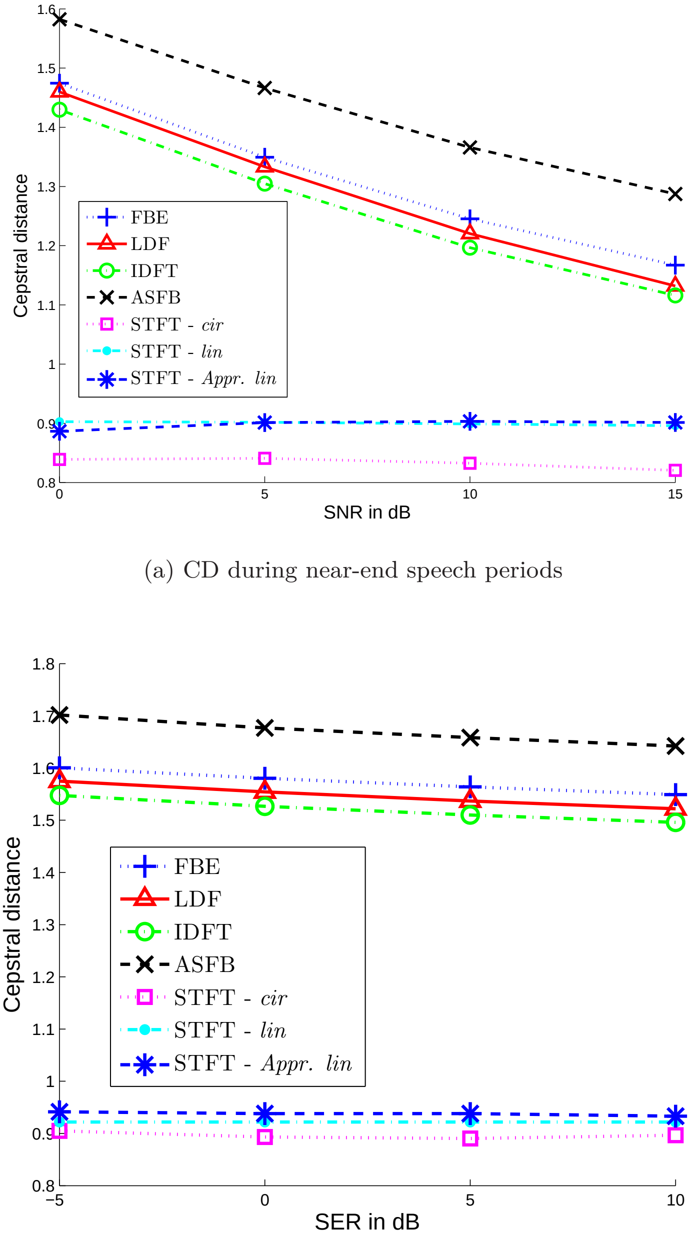
Figure 3.12(a) shows NA against SNR. The NA curves show that the STFT *Appr. lin* approach achieves the best performance in terms of noise reduction. The STFT *lin* and STFT *cir* methods achieve the worst performance but the gap between the different filtering approaches is small. In general, all the different filtering approaches are equivalent in terms of noise reduction: the differences between NA curves is less than 2dB. Lastly, it is of interest to note that the ranking of the FBE, LDF and ASFB performance is consistent with that reported in [Löllmann and Vary, 2007].

Figure 3.12(b) shows ERLE against SER. Here there is a clear gap between the amount of echo suppression achieved when gains are computed in the frequency domain compared to when it is computed in the subband domain. The STFT *lin*, STFT *cir* and STFT *Appr. lin* achieve the same amount of ERLE. The truncation of the interpolation function does not significantly impact the amount of echo suppressed.

The ASFB, FBE, LDF and IDFT approaches are equivalent in terms of echo suppression. We nevertheless observe that ERLE curves measured on clean speech signals (no additive noise) showed that the IDFT method achieves slightly less echo suppression compared to the ASFB, FBE and LDF method. In absence of noise, due to spectral structure of speech signals the attenuation difference between consecutive subbands can be very large. The more this attenuation difference increases, the more the effective frequency response of the IDFT filter has large variations (Gibbs phenomenon) between consecutive subbands. The results reported here address the problem of echo in the presence of noise, and this leads to a reduction in the attenuation difference between consecutive subbands and thus to better results for the IDFT approach.

Figure 3.13(a) shows cepstral distance against SNR during near-end speech only periods, i.e. distortions resulting from noise reduction. We see that the ASFB approach introduces the most distortions. The results regarding the ASFB, FBE and LDF are different from those presented in [Löllmann and Vary, 2007] in which the ASFB and FBE approaches were reported to produce speech of equivalent quality. Our explanation is that this difference is due to the analysis and synthesis filter banks which are defined differently. In [Löllmann and Vary, 2007], the analysis and synthesis filter banks used for the ASFB and FBE approaches are not the same whereas in our experiments they are. The frequency domain filtering methods clearly introduce the least distortions. Figure 3.13(b) shows CD against SNR during double talk periods. The ranking of the different filtering methods remains the same as in Figure 3.13(a). We note an increase of the CD values during double talk periods. This is justified by the fact that in double-talk periods both echo and noise reduction are active.

Informal listening tests reveal that near-end speech during double talk periods is distorted whereas no distortion is noticed during near-end only periods. Listening tests with weighted speech signals $\bar{s}(n)$ reveal the presence of small distortions of near-end speech during near-end speech only periods. These observations imply that echo processing brings more distortion than noise reduction no matter the filtering approach used. Distortion introduced by the noise reduction are not audible in processed speech signals due to the masking effect of residual noise present in processed speech signals. The distortion observed is mainly crackling noises for signals processed by time domain filtering methods, STFT *cir* and STFT *Appr. lin*. As explained in [**Yemdji** et al., 2010a], the crackling observed with time domain filters comes from the fact that their frequency responses are smoother than that of the original spectral gains which are defined per sub-band. The crackling observed in the STFT *cir* and STFT *Appr. lin* methods are respectively due to time-domain aliasing and the truncated weighting function $\tilde{z}_{l,i}$. We also note the presence of musical noise (random spectral peaks of short duration) for signals processed by the STFT *lin* and ASFB methods. The differences between signals processed by the IDFT, LDF and FBE approaches were hardly audible. This confirms what might be expected on account of results illustrated in Figure 3.13.



(a) CD during near-end speech periods

(b) CD during double-talk periods

Figure 3.13: Cepstral distance

3.4.3 Synthesis

The comparison of the frequency domain and subband domain related filtering methods shows that all filtering methods are equivalent in terms of NA. ERLE curves reported show that frequency domain filtering methods are slightly better than subband domain filtering methods in terms of ERLE. Lastly, plots of the CD shows that the frequency domain approach introduces less distortions than the subband domain related filtering methods. While the frequency domain filtering methods and subband filtering introduce some musical noise, the time domain filtering methods introduce some crackling noise in the output signal.

In the remainder of this thesis, we focus on the problem of echo and consider that there is no noise in the system. Experiments show that frequency domain filtering methods achieve good performance in terms perturbation of attenuation and useful speech distortion. Linear convolution is preferred for its mathematical exactness and will be the only filtering methods used in the remainder of this thesis.

3.5 Conclusion

In this chapter we present two different interpretations of the STFT: the filter bank and the DFT. Each interpretation leads to different filtering methods.

This chapter reports a side-by-side comparison of these filtering methods based on a combined NR and echo postfiltering algorithm. Results show that frequency domain filtering methods achieve more echo suppression than subband methods. It is mainly of interest to note that the proposed filtering method (that using the truncated weighting function) is a good alternative to linear and circular convolution.

3.A Time domain aliasing in circular convolution

3.A.1 Proof of aliasing in circular convolution

Let us consider two vectors $a(n)$ and $b(n)$ of length M. $A(i)$ and $B(i)$ are the Fourier transform of $a(n)$ and $b(n)$ respectively. The convolution of $a(n)$ and $b(n)$ in the time domain outputs $2M-1$ samples while the bin-by-bin multiplication of $A(i)$ and $B(i)$ outputs M samples.

Let us denote $C_1(n,i)$ the result of the bin-by-bin multiplication of $A(i)$ and $B(i)$

$$C_1(i) = A(i) \cdot B(i), \tag{3.35}$$

Its inverse Fourier transform corresponds to

$$c_1(n) = \frac{1}{M} \sum_{i=0}^{M-1} C_1(i) \, \exp\left(2\pi j \frac{ni}{M}\right) \tag{3.36}$$

$$c_1(n) = \frac{1}{M} \sum_{i=0}^{M-1} A(i) \cdot B(i) \, \exp\left(2\pi j \frac{ni}{M}\right) \tag{3.37}$$

By introducing the definition of $A(i)$ in Equation 3.37, we obtain the following

$$c_1(n) = \frac{1}{M} \sum_{i=0}^{M-1} \left[\sum_{m=0}^{M-1} a(m) \mathrm{exp}\left(-2\pi j \frac{mi}{M}\right)\right] \cdot B(i) \, \exp\left(2\pi j \frac{ni}{M}\right) \tag{3.38}$$

$$c_1(n) = \sum_{m=0}^{M-1} a(m) \left[\frac{1}{M} \sum_{i=0}^{M-1} B(i) \mathrm{exp}\left(2\pi j \frac{(n-m)i}{M}\right)\right] \tag{3.39}$$

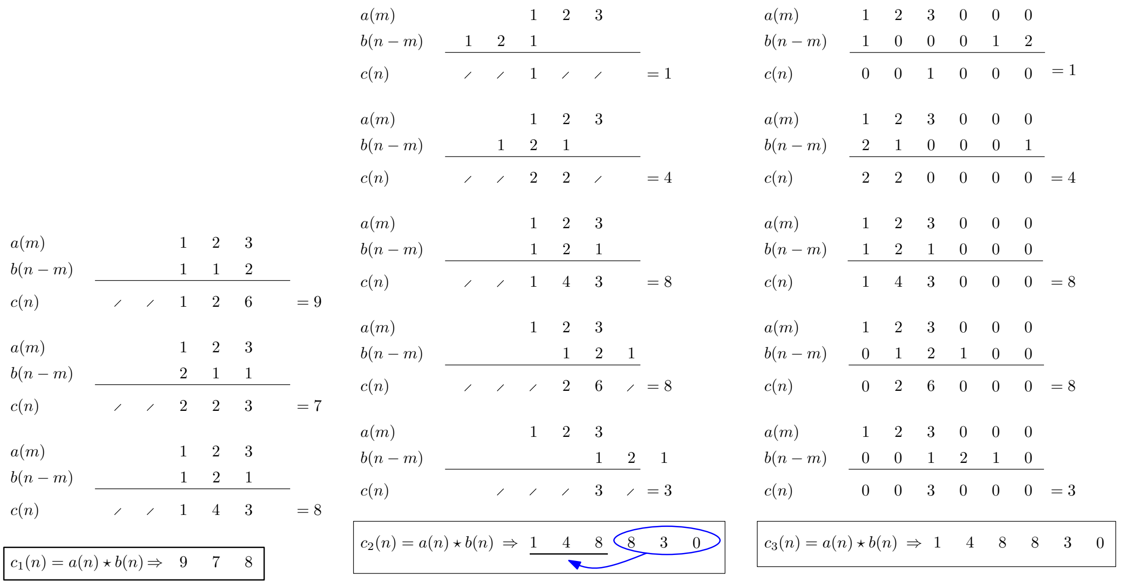
$$c_1(n) = \sum_{m=0}^{M-1} a(m) b\left((n-m)_M\right) = a(n) \star b\left((n)_M\right) \tag{3.40}$$

where $b\left((n)_M\right)$ denotes the circularly shifted version of $b(n)$. Equation 3.40 shows that the bin-by-bin multiplication of $A(i)$ and $B(i)$ is equal to the convolution of $a(n)$ with a circular shifted version of $b(n)$. In consequence $c_1(n)$ does not corresponds to the time domain convolution of $a(n)$ and $b(n)$. The circular shift that appears in Equation 3.40 is the source of time domain aliasing and can be source of distortion in speech processing.

3.A.2 Illustration of time domain aliasing in circular convolution

Time domain aliasing due to circular convolution is illustrated in Figure 3.14(a). The example in Figure 3.14 illustrates the convolution of two vectors $a(n) = [1\ 2\ 3]$ and $b(n) = [1\ 2\ 1]$. Ideally, the convolution of a and b should output c_2 (see Figure 3.14(b)). Instead, we see that the output of the circular convolution c_1 is a summed version of that of the standard time domain convolution Figure 3.14(b). The mismatch between c_1 and c_2 is due to time-domain aliasing.

Figure 3.14 (c) shows how zero-padding can be used to avoid time domain aliasing which occurs in circular convolution. We see that with appropriate zero-padding linear convolution can be achieved.



(a) Circular convolution

(b) Standard time domain convolution

(c) Linear convolution with circular shift

Figure 3.14: Circular and linear convolution in the domain

3.B Properties of the proposed interpolation function

In this section, we demonstrate that the interpolation function $z_{l,i}$ presented in Section 3.3.4 is normalized (sum for all i is equal to one) and that its real part is constant. We recall that in Equation 3.20 we defined $z_{l,i}$ as follows:

$$z_{l,i} = \frac{1}{M} \frac{2}{1 - \exp\left(\frac{2\pi}{2M} j \left(2i - l\right)\right)}. \tag{3.41}$$

The interpolation function $z_{l,i}$ is defined as a fractional complex number. In the following, we extract the expressions of the real and imaginary parts of $z_{l,i}$:

$$\begin{aligned}
z_{l,i} &= \frac{1}{M} \frac{2}{1 - \exp\left(\frac{2\pi}{2M} j \left(2i - l\right)\right)} & (3.42)\\
&= \frac{1}{M} \frac{2\Big(1 - \exp\left(-\frac{2\pi}{2M} j \left(2i - l\right)\right)\Big)}{\Big(1 - \exp\left(\frac{2\pi}{2M} j \left(2i - l\right)\right)\Big)\Big(1 - \exp\left(-\frac{2\pi}{2M} j \left(2i - l\right)\right)\Big)} & (3.43)\\
&= \frac{1}{M} \frac{2\Big(1 - \exp\left(-\frac{2\pi}{2M} j \left(2i - l\right)\right)\Big)}{1 - \exp\left(-\frac{2\pi}{2M} j \left(2i - l\right)\right) - \exp\left(\frac{2\pi}{2M} j \left(2i - l\right)\right)\Big) + 1} & (3.44)\\
&= \frac{1}{M} \frac{2\Big(1 - \exp\left(-\frac{2\pi}{2M} j \left(2i - l\right)\right)\Big)}{2 - 2\cos\left(\frac{2\pi}{2M} \left(2i - l\right)\right)} & (3.45)
\end{aligned}$$

$$= \frac{1}{M} \frac{1 - \cos\left(\frac{2\pi}{2M}(2i-l)\right) + j\sin\left(\frac{2\pi}{2M}(2i-l)\right)}{1 - \cos\left(\frac{2\pi}{2M}(2i-l)\right)} \tag{3.46}$$

$$= \frac{1}{M} + j\frac{1}{M}\frac{sin\left(\frac{2\pi}{2M}(2i-l)\right)}{1 - \cos\left(\frac{2\pi}{2M}(2i-l)\right)} = \Re\left(z_{l,i}\right) + j\Im\left(z_{l,i}\right) \tag{3.47}$$

We see from Equation 3.47 that the real part the interpolation function $z_{l,i}$ is constant for all values of i. This means that for a given frame k, the spectral gains used for the linear convolution (i.e. $\tilde{W}(k, l = 2i + 1)$) all have the same real part and simply differ by their imaginary part. In our assessement of this interpolation function, we study the impact of the truncation of $z_{l,i}$. But one can also imagine to apply the truncation only to the imaginary part of $z_{l,i}$ since its real part is constant.

Equation 3.47 shows that a given frame index, the interpolation function $z_{l,i}$ (used in the computation of $\tilde{W}(k, l = 2i + 1)$) does not equally weight the spectral gains $W(k, i)$. It is of interest to know whether for a given frame k, the sum of $z_{l,i}$ for all i (since interpolation is done along i) is normalized (i.e. equal to one.):

$$\sum_{i=0}^{M-1} z_{l,i} = \sum_{i=0}^{M-1} \Re\left(z_{l,i}\right) + j\Im\left(z_{l,i}\right) \tag{3.48}$$

$$= \sum_{i=0}^{M-1} \frac{1}{M} + j\frac{1}{M}\sum_{i=0}^{M-1} \frac{sin\left(\frac{2\pi}{2M}(2i-l)\right)}{1 - \cos\left(\frac{2\pi}{2M}(2i-l)\right)} \tag{3.49}$$

$$= 1 + j\frac{1}{M}\sum_{i=0}^{M-1} \frac{sin\left(\frac{2\pi}{2M}(2i-l)\right)}{1 - \cos\left(\frac{2\pi}{2M}(2i-l)\right)}. \tag{3.50}$$

For simplicity let us define α as follows

$$\alpha = \frac{2\pi}{2M}(2i-l) \tag{3.51}$$

and denote the term under the summation as follows

$$f(\alpha) = \frac{sin\alpha}{1 - cos\alpha}. \tag{3.52}$$

For a given value of l, α spans from $\frac{-\pi l}{M}$ to $2\pi - \frac{\pi l}{M}$ which is an interval of length 2π. By symmetry of the cos and sin functions, the summation of $f(\alpha)$ over an interval of length 2π is constant whether the summation interval spans from $\frac{-\pi l}{M}$ to $2\pi - \frac{\pi l}{M}$ or from 0 to 2π. If we consider $f(\alpha)$ for alpha ranging from 0 to 2π, we observe that it has the following symmetry $f(\pi - \alpha) = -f(\pi + \alpha)$. Therefore the summation of $f(\alpha)$ over this interval is null. Combining this result with Equation 3.50, we see that the sum of $z_{l,i}$ for all i is equal to 1.

Chapter 4

Synchronized adaptive echo cancellation and echo postfiltering

Contents

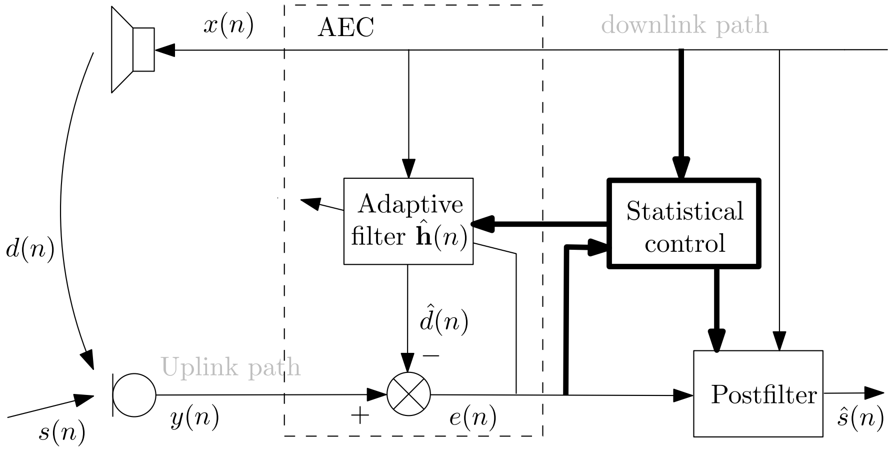


Figure 4.1: System overview

The previous chapter presented efficient filtering approaches to perturbations suppression for subband or frequency domain postfiltering. The postfilter used aimed at suppressing both residual echo and noise. In this chapter, we focus on the echo problem.

We explained in Chapter 2 that echo control systems are composed of adaptive filtering followed by residual echo suppression. Recent studies to improve single microphone echo processing performance focus on synchronizing both modules - synchronization which bounds the AEC and residual echo suppression to operate in the same frequency or subband domain. In this chapter, we introduce a cross-domain approach to synchronized echo control. Based on the similarities between our synchronization approach and the system in [Enzner and Vary, 2006], a new variable stepsize for the AEC is also introduced.

This Chapter is organized as follows. Section 4.1 presents the echo processing scheme of interest in this chapter. Our approach to synchronize the AEC and echo postfiltering is introduced in Section 4.2.

The cross-domain synchronization method presented in this chapter has been published in [**Yemdji** et al., 2012b].

4.1 System overview

Figure 4.1 shows an overview of the synchronized echo cancellation system of interest in this chapter. The microphone signal $y(n)$ is the sum of the near-end signal $s(n)$ and the echo signal $d(n)$ which is obtained by the convolution of the loudspeaker signal $x(n)$ with the acoustic path $\boldsymbol{h}(n)$. An adaptive filter is used to generate an estimate of the echo signal $\hat{d}(n)$ which is subtracted from the microphone signal to obtain the error signal $e(n)$. The error signal is composed of residual echo $\tilde{d}(n)$ and, possibly, of near-end speech $s(n)$. The postfilter aims to suppress the residual echo. In addition to conventional feedback used by the adaptive filter, an additional level of statistical control is applied to synchronize the adaptive filter and echo postfilter. The following details the investigated adaptive filter and echo postfilter.

4.1.1 Adaptive echo cancellation

The adaptive filter is based upon a normalized least mean square (NLMS) algorithm where the acoustic path estimate $\hat{\boldsymbol{h}}(n)$ and its optimum stepsize $\mu(n)$ are expressed as follows [Haykin, 2002]:

$$\hat{\boldsymbol{h}}(n+1) = \hat{\boldsymbol{h}}(n) + \frac{\mu(n)}{\mathbf{x}(n)^T \cdot \mathbf{x}(n)} \cdot \mathbf{x}(n) \cdot e(n) \quad \text{and} \tag{4.1}$$

$$\mu(n) = \frac{E\{\tilde{d}^2(n)\}}{E\{e^2(n)\}}, \tag{4.2}$$

where $\mathbf{x}(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-L+1) \end{bmatrix}^T$ is the loudspeaker signal, L is the length of the adaptive filter and $E\{.\}$ represents statistical expectation. The computation of the variable stepsize requires knowledge of the residual echo power $E\{\tilde{d}^2(n)\}$ which is not directly measurable. Instead, it is approximated through the system distance as follows [Hänsler and Schmidt, 2004, Haykin, 2002]:

$$E\{\tilde{d}^2(n)\} = E\{x^2(n)\} \cdot \|\boldsymbol{\beta}_L(n)\|^2 \tag{4.3}$$

where $\boldsymbol{\beta}_L(n)$ is the system mismatch i.e. the error between the real acoustic path $\boldsymbol{h}(n)$ and its estimate $\hat{\boldsymbol{h}}(n)$. The value $\|\boldsymbol{\beta}_L(n)\|^2$ is referred to as the system distance [Haykin, 2002] and its computation is described in Section 4.2.

4.1.2 Echo postfiltering

The postfilter consists of frequency domain processing with filtering through linear convolution in the frequency domain [Oppenheim and Schafer, 1999, **Yemdji** et al., 2011]. Prior to frequency gain computation, the postfilter input signals $x(n)$ and $e(n)$ are converted into frequency domain signals $x(k,i)$ and $e(k,i)$, respectively through DFT comprising overlap (see Section 3.1.2). The i^{th} frequency signal $\hat{s}(k,i)$ is obtained through the multiplication of the gain $W(k,i)$ with $e(k,i)$. Conversion from time to frequency domain is performed on blocks of R samples through a fast Fourier transform with an overlap-add method [Oppenheim and Schafer, 1999].

For each frequency index i, the postfilter gains $W(k,i)$ are computed according to the Wiener rule as in Section 3.4.1.1:

$$W(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)}, \tag{4.4}$$

where $\xi(k,i)$ is the signal (near-end speech) to (residual) echo ratio (SER). As described in Section 3.4.1.1, the SER is estimated through the Ephraim and Malah approach. Its computation requires an estimate of the residual echo power which we implement as:

$$\hat{\Phi}^{\tilde{d}\tilde{d}}(k,i) = |\tilde{H}(k,i)|^2 \cdot \Phi^{xx}(k,i), \tag{4.5}$$

where $\hat{\Phi}^{\tilde{d}\tilde{d}}(k,i)$ is the residual echo power spectral density, $\hat{\Phi}^{xx}(k,i)$ is the loudspeaker power spectral density and $|\tilde{H}(k,i)|^2$ is the system mismatch power spectrum [Steinert et al., 2007]. The computation of $|\tilde{H}(k,i)|^2$ is described in Section 4.2.

4.2 System control

In this section, we present our approach to synchronize the AEC and the echo postfilter. The synchronization approach presented here operates with fullband AEC but can be adapted readily to operate with a subband AEC. Thus AEC and echo postfiltering are not constrained to operate in the same domain.

4.2.1 Synchronization approach

The architecture used here is inspired from existing synchronized approaches to echo control such as those in [Enzner and Vary, 2003, 2006, Steinert et al., 2007]. In such systems, the acoustic echo canceler is constrained to function in the frequency or subband domains. However, a comparative study shows that subband or frequency domain AECs are less robust to non-linearities than fullband AECs [Mossi et al., 2010]. The comparative assessment in [Mossi et al., 2010] of the behavior of AECs showed that frame based adaptive filters such as the frequency block LMS algorithm are less robust to non-linearities than sample-by-sample based AECs. Indeed, for regions where the loudspeaker signal is low, non-linear effects are negligible. Sample-by-sample based AEC will be able to estimate the echo path. Whereas for block-based AEC required a whole frame of low amplitude loudspeaker signal to benefit from the same effect.

Our approach to synchronization is based upon the correspondence between the system mismatch $\boldsymbol{\beta}_L(n)$ and its spectrum $\tilde{H}(k,i)$ which are defined as follows

$$\boldsymbol{\beta}_L(n) = \boldsymbol{h}(n) - \hat{\boldsymbol{h}}(n) \quad \text{and} \quad \tilde{H}(k,i) = H(k,i) - \hat{H}(k,i), \tag{4.6}$$

where $H(k,i)$ and $\hat{H}(k,i)$ are the discrete Fourier transforms (DFT) of $\boldsymbol{h}(n)$ and $\hat{\boldsymbol{h}}(n)$ respectively. From Equation 4.6, we note that $\tilde{H}(k,i)$ is the DFT of $\boldsymbol{\beta}_L(n)$. According to Parseval's equality [Proakis and Manolakis, 1996], we can write the following:

$$\|\boldsymbol{\beta}_L(n)\|^2 = \frac{1}{M}\sum_{i=0}^{M-1} |\tilde{H}(k,i)|^2, \tag{4.7}$$

with the assumption that n is a multiple of the blocksize R. Equation 4.7 highlights the relationship between the NLMS algorithm and the echo postfilter. This relationship can be used in two different ways:

- The estimate of the system mismatch $\boldsymbol{\beta}_L(n)$ can be used to compute its norm $\|\boldsymbol{\beta}_L(n)\|^2$ and its power spectrum $|\tilde{H}(k,i)|^2$. This solution is impractical because the misalignment vector $\boldsymbol{\beta}(n)$ cannot be estimated reliably. The estimation of $\boldsymbol{\beta}_L(n)$ requires correlation computation [Haykin, 2002] which is highly computationally demanding. Most real time systems estimate the system distance directly [Hänsler and Schmidt, 2004].
- Alternatively, $|\tilde{H}(k,i)|^2$ can be estimated and used to derive $\|\boldsymbol{\beta}_L(n)\|^2$ according to Equation 4.7. As most echo postfilters already require the computation of $|\tilde{H}(k,i)|^2$, we opted for this solution. In this case there is no additional computational requirement.

The system mismatch power spectrum $|\tilde{H}(k,i)|^2$ can be computed through the cross-correlation method [Steinert et al., 2007] according to:

$$|\tilde{H}(k,i)|^2 = \left|\frac{\Phi^{xe}(k,i)}{\Phi^{xx}(k,i)}\right|^2, \tag{4.8}$$

where $\Phi^{xe}(k,i)$ is the cross spectral density between $e(n)$ and $x(n)$. However, the postfilter is updated on a frame-by-frame basis whereas the AEC requires a sample-by-sample update. In between two measurements of the system mismatch power spectrum, the system distance is updated according to the following recursion [Claasen and Mecklenbrauker, 1981, Steinert et al., 2007]:

$$\|\boldsymbol{\beta}_L(n+1)\|^2 = \Big(1 - \frac{\mu(n)}{L}\Big) \cdot \|\boldsymbol{\beta}_L(n)\|^2. \tag{4.9}$$

A similar recursion can be found in the echo control system in [Enzner and Vary, 2006]. In the next section, we propose a novel recursion for the computation of the system distance based on the similarities between Equation 4.9 and the work in [Enzner and Vary, 2006].

4.2.2 Enhanced variable stepsize

The system in [Enzner and Vary, 2006] can be seen as a variable stepsize NLMS with an adaptive filter that has one tap per frequency bin. Therefore, $|\tilde{H}(k,i)|^2$ defines the system distance for each NLMS filter or frequency bin. In [Enzner and Vary, 2006], $|\tilde{H}(k,i)|^2$ is computed as:

$$|\tilde{H}(k+1,i)|^2 = A^2 \cdot [1 - \frac{R}{M}\mu(k,i) \cdot |x(k,i)|^2] \cdot |\tilde{H}(k,i)|^2 + (1-A^2) \cdot |\hat{H}(k,i)|^2 \tag{4.10}$$

$$|\tilde{H}(k+1,i)|^2 = A^2 \cdot [1 - \frac{R}{M}\mu(k,i) \cdot |x(k,i)|^2] \cdot |\tilde{H}(k,i)|^2 + |\Delta H(k,i)|^2 \tag{4.11}$$

where A is a constant which models variations of the acoustic path and A should be comprised between 0.9 and 0.999. Values of A closed to 1 model small acoustic path change and vice-versa. The static echo path modeling corresponds to $A = 1$. The 2^{nd} term of the summation $|\Delta H(k,i)|^2$ is equal to 0 for static echo path modeling. This means that the term $|\Delta H(k,i)|^2$ permits to account for accounts the variability of the echo path. To improve the performance of our system, we propose a new measure of the system distance.

According to Equation 4.7, a measure of the system distance over all frequency bins can be derived from Equation 4.11:

$$\|\boldsymbol{\beta}_L(n+1)\|^2 = \frac{1}{M}\sum_{i=0}^{M-1} |\tilde{H}(n+1,i)|^2 \tag{4.12}$$

$$\approx A^2 \cdot (1 - \frac{R}{M}\mu(n,i) \cdot |x(n,i)|^2)\|\boldsymbol{\beta}_M(n)\|^2 + (1-A^2) \cdot \frac{1}{M}\sum_{i=0}^{M-1} |\hat{H}(n,i)|^2 \tag{4.13}$$

$$\approx A^2 \cdot (1 - \frac{R}{M}\mu(n,i) \cdot |x(n,i)|^2)\|\boldsymbol{\beta}_M(n)\|^2 + (1-A^2) \cdot \|\hat{h}(n)\|^2 \tag{4.14}$$

$$\approx A^2 \cdot (1 - \frac{R}{M}\mu(n,i) \cdot |x(n,i)|^2)\|\boldsymbol{\beta}_M(n)\|^2 + \|\Delta \boldsymbol{h}(n)\|^2. \tag{4.15}$$

Equations 4.9 and 4.15 both defined the system distance in the form of a recursion except for the additional term $\|\Delta \boldsymbol{h}(n)\|^2$ which is not present in Equation 4.9. Based on Equation 4.15, we redefine the system distance in Equation 4.9 by adding a second term $\|\Delta \boldsymbol{h}(n)\|^2$ as follows:

$$\|\boldsymbol{\beta}_L(n+1)\|^2 = \Big(1 - \frac{\mu(n)}{L}\Big) \cdot \|\boldsymbol{\beta}_L(n)\|^2 + \|\Delta \boldsymbol{h}(n)\|^2, \tag{4.16}$$

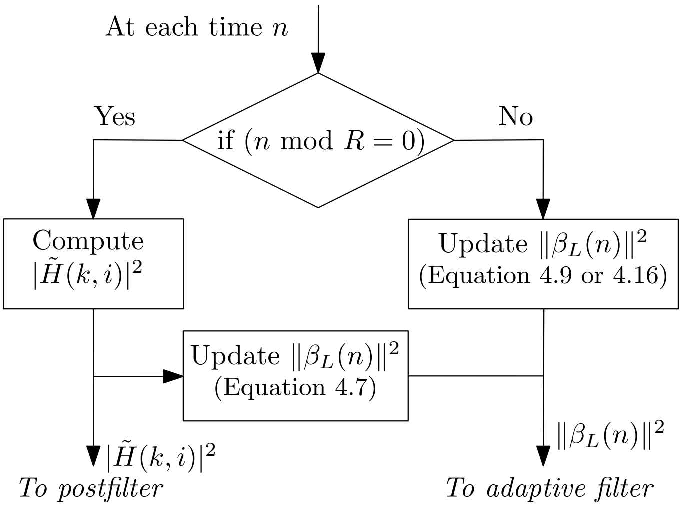

At each time n
if (n mod $R = 0$)
Yes
No
Compute $|\tilde{H}(k,i)|^2$
Update $\|\beta_L(n)\|^2$ (Equation 4.9 or 4.16)
Update $\|\beta_L(n)\|^2$ (Equation 4.7)
$|\tilde{H}(k,i)|^2$
To postfilter
$\|\beta_L(n)\|^2$
To adaptive filter

Figure 4.2: Statistical control diagram

where $\|\Delta\boldsymbol{h}(n)\|^2$ accounts for changes in the acoustic path. Equivalently, $\|\Delta\boldsymbol{h}(n)\|^2$ is computed according to:

$$\|\Delta\boldsymbol{h}(n)\|^2 = (1 - A^2) \cdot \|\hat{\boldsymbol{h}}(n)\|^2. \tag{4.17}$$

The $\|\Delta\boldsymbol{h}(n)\|^2$ is defined in Equation 4.17 as a positive quantity. Its use according to Equation 4.16 will impact on the behavior of the AEC. By comparing equations 4.9 and 4.15, we can state that the newly defined system distance will lead to values of $\mu(n)$ that are higher than those obtained with Equation 4.9.

4.2.3 Summary

The echo control scheme considered in this chapter is composed of an adaptive filter followed by an echo postfilter as illustrated in Figure 4.1. The novelty in the system presented here resides in the use of the statistical control module which we use to link the AEC and echo postfilter. The synchronization approach is summarized in Figure 4.2. It computes the system mismatch power spectrum according to Equation 4.8 on a frame-by-frame basis (i.e. when $n \mod R = 0$). $|\tilde{H}(k,i)|^2$ is used within the postfilter to update the spectral gains $W(k,i)$ according to Equations 4.4 and 4.5 and within the adaptive filter for the computation of the system distance according to Equation 4.7. During intervals in which the postfilter is not updated, the system distance is updated according to Equations 4.9 or 4.16. We note that Equation 4.9 is equal to Equation 4.16 for $A = 1$.

4.3 Experiments

In this section we assess the synchronized echo control system proposed above and compare its performance to approaches. Although this chapter reports synchronized echo control, AEC performance and echo postfilter performance are nonetheless assessed separately.

In Section 4.3.1 we present our experimental setup. Section 4.3.2 reports an analysis of the impact of the synchronization on the convergence of the AEC while in Section 4.3.3 we report the performance of the whole echo control system.

4.3.1 System setup

The proposed synchronized echo control systems are compared to 2 state-of-the-art systems. State-of-the-art systems considered are:

- The unsynchronized system composed of a NLMS with fixed stepsize ($\mu = 0.1$) followed by a postfilter. The postfilter is that describe in Section 4.1.2 meaning it is the same as that uses in our synchronized system except that it is not linked to AEC.
- The synchronized Kalman echo control system which we denote *Kalman AEC* [Enzner and Vary, 2006].

The first set of experiments presented in Section 4.3.2 assesses the impact of our synchronization approach on the AEC. The AEC part is assessed in terms of robustness to echo path changes and convergence time. We assess the interest of the proposed system distance measure (Equation 4.16) by comparing its performance to that of the system distance as defined in Equation 4.9. Both systems differ in terms of the value of A which we set to 0.99 as in [Enzner and Vary, 2006] within Equation 4.16. We also show the interest of the synchronization between the AEC and the postfilter by assessing the performance of the AEC when linked with the postfilter to the performance of the AEC when used independently. The second set of experiments reports performances of the whole echo control scheme (i.e. AEC followed by postfiltering.) Table 4.1 summarizes the different echo control systems considered.

For the systems considered, the number of frequency bins M is set to 256 while the framesize R is set to 128. Filtering of the residual echo in the postfilter takes place through linear convolution in frequency domain (see Section 3.3.2). Lastly, the length of the adaptive filter is set to 256.

All simulations reported here were performed with speech signals. Microphone speech signals contain an echo-only interval followed by a double-talk interval. The echo-only interval is long enough (8s) so that each AEC algorithm converges. The double-talk interval is used to assess the impact of near-end speech on both the AEC and postfilter. The echo signals are generated by convolving the loudspeaker signal with an acoustic path response. Four different acoustic path responses are used; they were all measured with real mobile terminals in an office environment. The resulting database of speech signals has SERs ranging from -5 dB to 10 dB with the near-end speech level set to -26 dB. Speech signal levels are set through the ITU-T speech voltmeter [ITU-T, 1993].

Performance is assessed in terms of echo return loss enhancement (ERLE), cepstral distance and informal listening tests. While the ERLE is used to assess the amount of echo suppression during echo-only intervals, the cepstral distance is used to assess the amount of distortion introduced by postfiltering during double-talk intervals.

4.3.2 Convergence of the AEC

This section reports the performance of the AEC part alone. We demonstrate the interest of our synchronization method and analyze its performance in case of echo path change.

4.3.2.1 Interest of the synchronization

In this section, we analyze the impact of the proposed synchronization module on the convergence of AEC. Experiments reported in this section are based on microphone signals

	System distance	Synchronization
Unsync. A = 1	NLMS algorithm with SD according to Equation 4.9	No
Unsync. A = 0.99	NLMS algorithm with SD according to Equation 4.16	No
Sync. A = 1	NLMS algorithm with SD according to Equation 4.9	Yes
Sync. A = 0.99	NLMS algorithm with SD according to Equation 4.16	Yes
Fixed stepsize	NLMS algorithm with fixed stepsize	No
Kalman AEC	Kalman AEC from [Enzner and Vary, 2006]	Yes

Table 4.1: Summary of algorithms tested

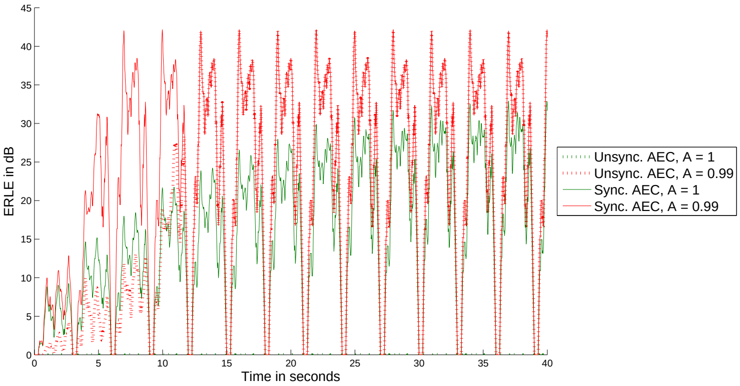


Figure 4.3: Impact of the synchronization on the AEC

containing echo-only. To do so, we compare the performance of the AEC with and without the synchronization with the postfilter (synchronization refers to Equation 4.7).

Figure 4.3 reports the ERLE curve for the synchronized and unsynchronized version of the system proposed in this chapter. We see that when the AEC does not receive any feedback from the postfilter and that we compute the system distance according to Equation 4.9, the AEC is not effective in canceling the echo. Whereas when the system distance is computed as we propose (Equation 4.16), the AEC achieves up to 40 dB echo attenuation. For both estimations of the system distance, the synchronization improves the performance of the AEC. When using Equation 4.16, the synchronization permits to significantly reduce the convergence time of the AEC. When the system distance is computed according to 4.16, the synchronization permits to increase the amount of echo suppressed by the AEC. In fact, the synchronization with the postfilter leads to high stepsize values and forces the AEC to adapt.

This experiment shows that the synchronization permits to improve the performance of AEC. We also see that the proposed system distance estimate also leads to better echo cancellation. In the remaining of this Chapter, only the synchronized version *Sync. A =1* and *Sync. A =0.99* of the proposed system is used. In other terms, system distance is computed according to Equation 4.9 or 4.16 and is always synchronized with the postfilter (Equation 4.7).

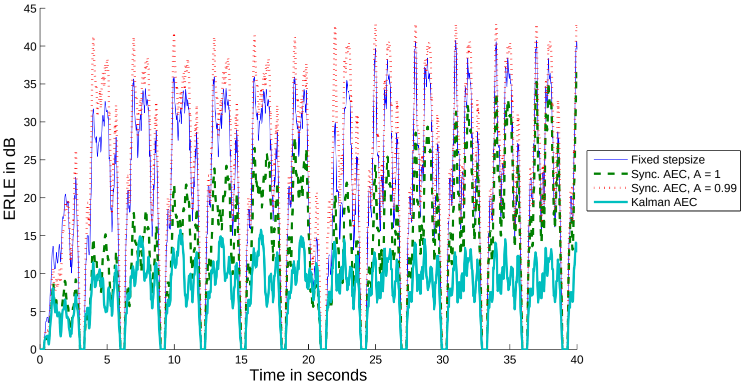


Figure 4.4: ERLE against time for AEC. An abrupt echo path change occurs at time $t = 20s$.

4.3.2.2 Robustness to echo path changes

In this section, we assess the robustness of the different AECs considered to echo path changes. Both abrupt and slowly varying echo path changes are considered.

Figure 4.4 illustrates the convergence of the AEC in case of an abrupt echo path change. At time $t = 20s$, there is an abrupt echo path change. The curves show that the *Sync. A = 0.99* system converges faster than the *Sync. A = 1* system. Its rapidity is due to the term $\|\Delta \boldsymbol{h}(n)\|^2$ which is used in the *Sync. A = 0.99* system and not in the *Sync. A = 1* system. $\|\Delta \boldsymbol{h}(n)\|^2$ leads to higher stepsize values and therefore to faster convergence. The *Sync. A = 0.99* system also achieves more echo suppression than the *Sync. A = 1* system.

Performance of the NLMS based AECs is severely affected by the abrupt echo path change. We note an important decrease in their ERLE values. Although, the Kalman system achieves the least ERLE, it is the most robust to the abrupt echo path change at time $t = 20s$. The poor performance of the Kalman system might be due to the model mismatch (except for the echo path change at $t = 20s$, the echo path is static) and to the fact the Kalman system operates in the frequency domain. The Kalman system is updated in the frequency domain on a frame-by-frame basis whereas the new approach is updated sample-by-sample.

Figure 4.5 shows the ERLE along time for a slowing varying echo path. The echo path variations are generated according the Markov model as in [Enzner and Vary, 2006, Malik and Enzner, 2008]. We see in Figure 4.5 that the NLMS with fixed stepsize and the *Sync. A = 0.99* achieves the most ERLE. We also see that the Kalman system converges even faster than in Figure 4.4 where the echo paths were static. This is because the Kalman system is design for time varying echo paths as it is the case here. Last, we note that the *Sync. A = 1* requires a long time to converge: this once more shows the benefit of including the term $\|\Delta \boldsymbol{h}(n)\|^2$.

4.3.3 Assessment of the global echo control scheme

Figure 4.6 shows the mean ERLE against SER for the four different AEC implementations considered. The (*Sync. A = 0.99*) system achieves the best performance in terms of ERLE. The proposed system (*Sync. A = 0.99*) gives better performance than the system

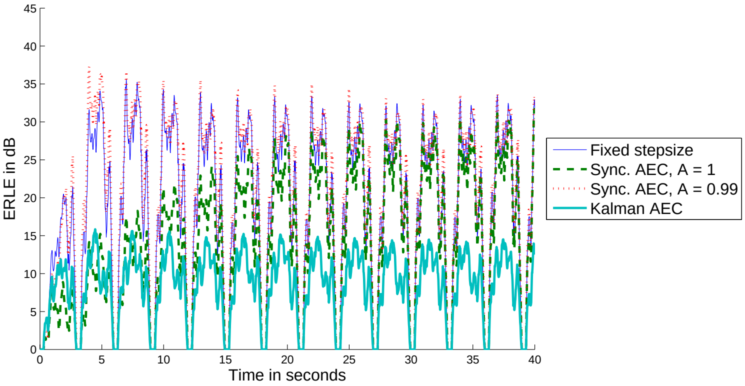


Figure 4.5: ERLE against time for AEC. Echo path is time varying.

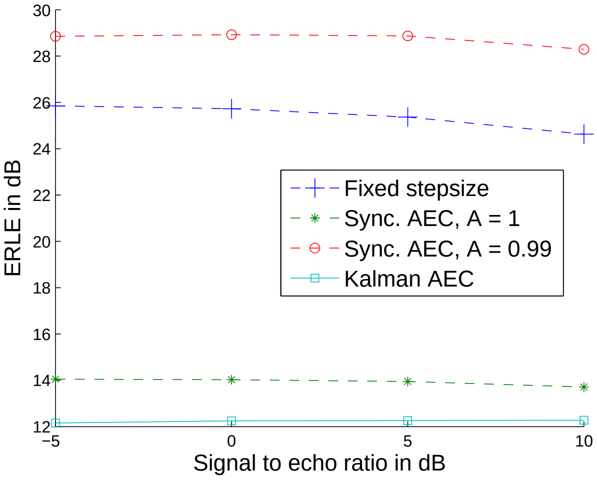


Figure 4.6: Average ERLE against SER for AEC only during echo-only intervals

distance (*Sync. A = 1*): more than 10 dB difference in ERLE across the full range of SERs. Nevertheless, the new system distance approach *Sync. A = 1* gives marginally better echo suppression than the Kalman AEC algorithm. This might be because the Kalman system is updated in the frequency domain on a frame-by-frame basis whereas the new approach is updated sample-by-sample.

Figure 4.7 illustrates the total amount of echo suppression achieved through combined AEC and postfiltering. The unsynchronized system and the Kalman system achieve the most echo suppression. The system with *Sync. A = 0.99* achieves slightly less echo suppression than the Kalman echo control system. This loss of performance can be attributed to the system mismatch power spectrum function estimate which is not the same in each postfilter. *Sync. A = 1* achieves the worst performance in terms of ERLE: this is attributed to poor AEC performance.

Figure 4.8 shows the mean cepstral distance against SER for the four different systems. The cepstral distance is measured at the output of the postfilter during double-talk periods. We observe that the system with fixed stepsize brings the most distortion. The Kalman echo control system brings the least distortion. Although the new synchronized

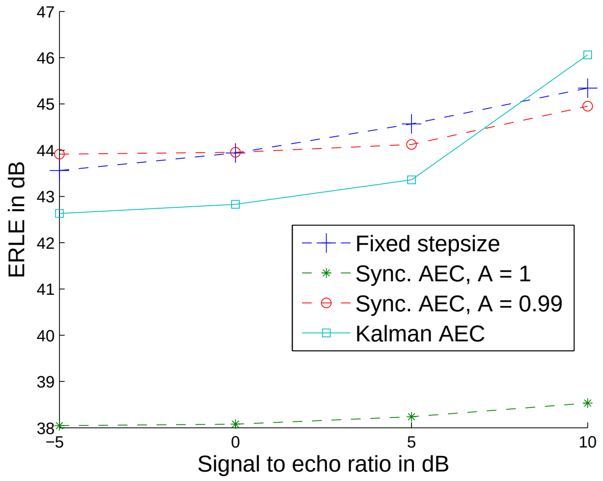


Figure 4.7: Average ERLE achieved through the AEC and the echo postfilter

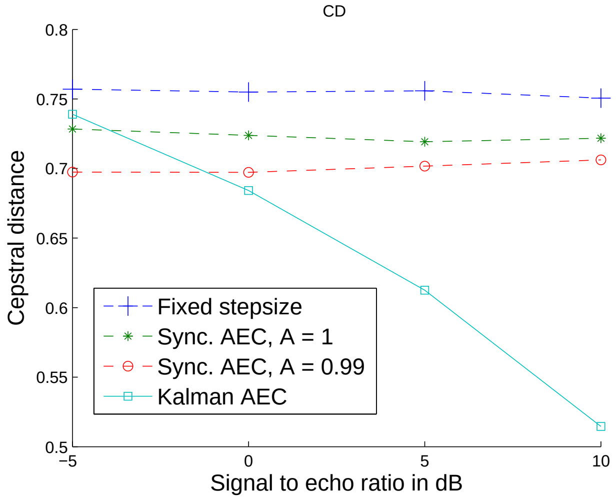


Figure 4.8: Average cepstral distance against SER for the global echo control system during double-talk periods

approaches introduce more distortion than the Kalman system, their levels of distortion remain low compared to that of the unsynchronized system. Nevertheless, the postfilter with *Sync. A =1* introduces slightly more distortion than the postfilter with *Sync. A = 0.99*. This comes from the fact that the AEC from *Sync. A = 1* achieves less echo suppression and thus places an increased demand on the postfilter than with AEC *Sync. A = 0.99*. Moreover, the complete echo control system *Sync. A = 1* achieves less echo suppression than the *Sync. A = 0.99* system (see Figure 4.6). The postfilter *Sync. A = 1* can be tuned in order to achieve as much echo suppression as *Sync. A = 0.99* but this results in increased distortion during double-talk intervals.

Informal listening tests reveal the presence of musical noise in signals at the output of the postfilter for both the proposed and the Kalman echo control systems. In addition to musical noise, signals processed by Kalman echo control sometimes contain crackling noise which was sometimes perceived as annoying. In signals processed by *Sync. A = 1*, echo is sometimes still audible whereas in signals processed by *Sync. A = 0.99* echo is inaudible.

4.4 Conclusion

This chapter presents the first cross-domain approach to synchronized acoustic echo cancellation and echo postfiltering. The proposed approach is based on the link between the system distance and the system mismatch power spectrum. A new system distance estimate is also introduced and assessed. The performance of the new synchronized echo control system is compared to synchronized Kalman echo control system and to an unsynchronized approach.

Our approach yields a reduction in distortion compared to the unsynchronized echo control system. The proposed system is robust to echo path changes and is stable during intervals of double-talk. The new system distance estimate delivers significantly improved echo suppression and rapid AEC convergence while preserving a reduced level of distortion during double-talk intervals compared to the standard system distance.

Part II

Dual microphone echo processing

Introduction

Until now, we have considered the echo problem for single microphone (SM) terminals. Typical SM echo processing schemes are composed of adaptive filtering followed by residual echo suppression. AEC can be achieved through various existing approaches such as LMS, NLMS, RLS or subband adaptive filter. As presented in Chapter 2, these adaptive filters each have relative merits and disadvantages but all lead to residual echo. Postfilters are required to suppress residual echo but they sometimes result in strong distortion of near-end speech.

Alternative approaches to improved echo cancellation are based on multi-microphone systems [Kellermann, 1997, Reuven et al., 2007a]. Multi-microphone echo cancellation approaches are based on beamforming techniques and have been shown to outperform SM approaches. Existing beamforming techniques typically require 4 to 10 microphones [Myllyla and Hamalainen, 2008, Reuven et al., 2007b] spaced by a distance of about 10cm up to 1 meter.

Mobile devices have traditionally been equipped with one microphone. Given the reduced size of mobile device, it would be very difficult to efficiently place 4 microphones on a device. Moreover, increasing the number of microphone on mobile device might result in a significant increase of the price of the device. In consequence, the device might not be competitive because of its price.

Nevertheless, to gain advantage of the potential of multi-microphone architecture, more and more dual microphone mobile phones can be found on the market. There is therefore a necessity to design dual microphone (DM) speech enhancement algorithms for mobile devices.

In contrast to DM noise reduction [Dörbecker and Ernst, 1996, Jeub et al., 2012], DM echo control has not received much attention [Guo et al., 2011, Jeannes et al., 2001]. The echo control systems in [Guo et al., 2011, Jeannes et al., 2001] both use adaptive filtering: one per microphone. The use of 2 AECs is a computationally prohibitive point for mobile devices. In [Jeannes et al., 2001] a combination of several postfilters is used to process residual echo. In [Guo et al., 2011], a beamformer is placed after the AECs to steer the error signals towards the direction of the desired speech signal $s(n)$. This means that in [Guo et al., 2011] residual echo is considered as an interfering signal and its suppression is achieved by the beamformer. In both cases (combination of postfilters or use of beamforming), the resulting DM echo cancellation system is still very computationally demanding for a mobile device.

In this 2^{nd} part of the thesis, we report our contributions regarding DM echo cancellation. First, we study the echo problem based on some recordings with mock-up and real DM mobile devices. These recordings are later on used to propose methods to improve echo cancellation in DM terminals. Proposed methods uses a conventional AEC followed by a DM postfilter. Experiments show that proposed methods outperforms SM echo control systems.

This part is organized as follows. In Chapter 5, an analysis of the echo problem for DM is presented. A DM echo postfilter and double-talk detector (DTD) are also presented and showed to be efficient for a certain arrangement of the transducers on the phone. In Chapter 6 another approach to DM echo postfilter is presented. This method is showed to be efficient for all transducer arrangement and is extendable to non-linear echo suppression.

Chapter 5

Echo cancellation for dual channel terminals

Contents

In this chapter we focus on the echo problem for dual-microphone terminals. Such an architecture is typical of some mobile terminals or tablets in today's market. For instance, the *iPhone 4*, *Google Nexus One* and *Samsung S series* are dual-microphone mobile terminals.

We propose a study of the echo problem for dual-microphone terminals. The proposed analysis is based on both handset and handsfree scenarios. Its follows from this analysis that two main features can be used for the purpose of echo cancellation: the level difference and the correlation between the microphone signals.

Both features require to define a novel dual-microphone echo processing scheme. The proposed scheme is very similar to that used in single microphone (SM) terminals and is composed of AEC followed by echo postfiltering. The similarity of the proposed scheme with the SM scheme is deliberate and is discussed further in this chapter. This chapter solely focuses on the use of level difference while correlation based methods will be presented in Chapter 6.

The level difference between the microphone signals is exploited in two different ways. As a first step, we simply use the level difference to introduce a novel frequency domain based double-talk detector (DTD). The proposed DTD is used as post-processing to enhance the SM echo postfilter used in previous chapters. As a second step, the level difference is used to introduce a new implementation of the well-known Wiener echo postfilter. Unlike most Wiener echo postfilters, that proposed here does not require an explicit estimate of the residual echo power spectral density.

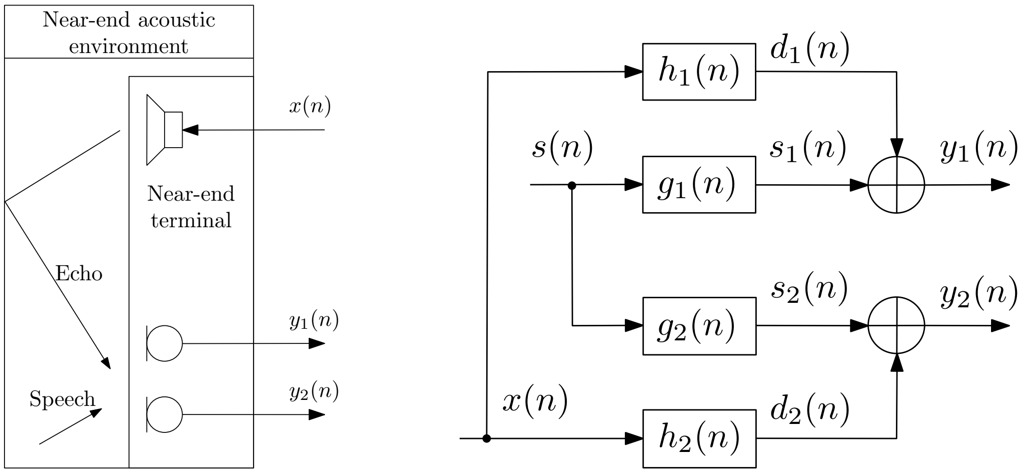
This chapter is organized as follows. Section 5.1 presents our model of the echo problem and the recordings performed with DM devices. Section 5.2 presents the proposed DM echo processing scheme. Our level difference based DTD and echo postfilter are presented in sections 5.3 and 5.4 respectively. Section 5.5 deals about experimental assessment of our DM approaches including a comparison with a baseline DM system. Lastly our conclusions are reported in Section 5.6.

The signal analysis and the proposed echo processing scheme presented in this chapter have been partly published in [**Yemdji** et al., 2012c]. The DTD and proposed gain rule have been patented [**Yemdji** et al., 2013].

5.1 Echo problem in dual channel terminals

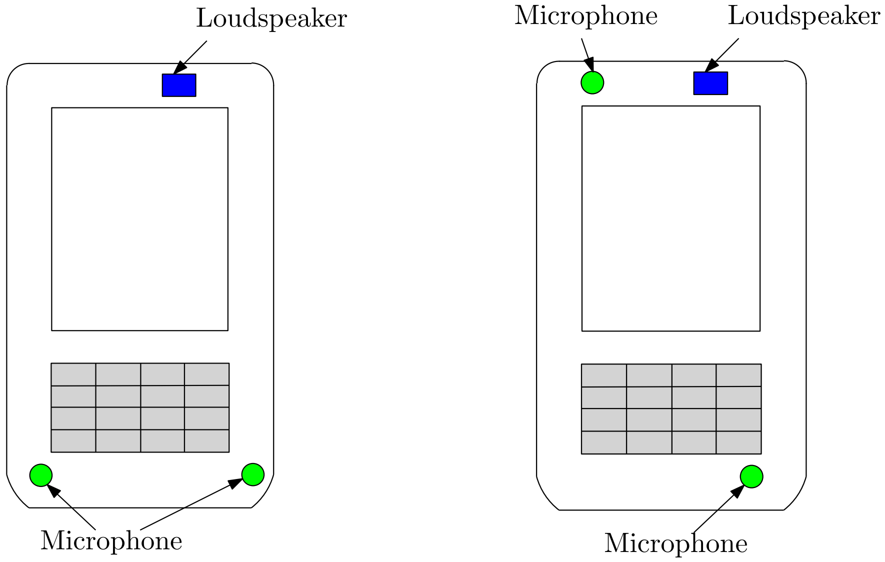
As illustrated in Figure 5.1 (a), we assume a terminal equipped with one loudspeaker and two microphones. Later, we consider one of the microphone signals to be the primary microphone signal and the other to be secondary. The primary and secondary microphones are denoted $y_1(n)$ and $y_2(n)$ respectively. Note that the positions of the microphones on Figure 5.1 (a) are not representative of their actual position on the terminal. These positions are simply used as illustration of the general set-up.

Figure 5.2 shows two examples of transducer configuration for mobile terminals. In the bottom-bottom configuration (Figure 5.2 (a)), the microphones are both placed at the bottom of the phone and are approximately equidistant from the loudspeaker. In the bottom-top configuration (Figure 5.2 (b)), the microphones are placed such that one is close to the loudspeaker whereas the other is relatively further away. We can cite as example the *iPhone 4* and *Samsung Galaxy S2* that use two microphones which are placed in bottom-top configuration. This configuration is the most widespread among terminals of today's market since it leads to features like the level difference and coherence between the



(a) Dual-microphone terminal (b) Dual-microphone terminal signal model

Figure 5.1: Illustration of the echo problem in a dual-microphone terminal



(a) Microphones in bottom-bottom configuration (b) Microphones in bottom-top configuration

Figure 5.2: Example of mobile device with different microphone positions

microphone signals which are highly exploited to improve performance of noise reduction algorithms [Jeub et al., 2012].

The examples of transducer positions in Figure 5.2 show the necessity to accounts for the different acoustic paths defined by the acoustic sources present in our system (near-end speaker and loudspeaker). Our study of DM echo processing methods aims to propose solutions that can be easily implemented on real devices. Since most terminal that are in the market today are bottom-top, our study will focus on this configuration.

5.1.1 Signal model

As depicted in Figure 5.1 (a), the far-end speaker voice is played by the loudspeaker to the near-end speaker. Part of this loudspeaker signal is reflected in the near-end environment and is recorded by both microphones [Hänsler and Schmidt, 2004, chap. 3]. The signals $d_1(n)$ and $d_2(n)$ represent the echo signal at the primary and secondary microphone respectively:

$$d_j(n) = h_j(n) * x(n) \tag{5.1}$$

where $j \in \{1, 2\}$, $x(n)$ stands for the loudspeaker signal and $h_j(n)$ denotes the acoustic echo path between the loudspeaker and the microphone j.

The microphones also record the speech signal from the near-end speaker and eventually the background noise. Similarly as for the echo, the speech from the near-end speaker is reflected in the surrounding environment before being recorded by the microphones [Jeub et al., 2011, Reuven et al., 2007a]. The signals $s_1(n)$ and $s_2(n)$ represent the near-end speech signal picked by the primary and secondary microphone respectively:

$$s_j(n) = g_j(n) * s(n) \tag{5.2}$$

where $j \in \{1, 2\}$, $s(n)$ denotes the near-end speech signal and $g_j(n)$ denotes the acoustic path between the near-end speaker's mouth and the microphone.

Given these explanations, the signal model of the dual microphone (DM) echo problem can be schematized as shown in Figure 5.1 (b). The resulting primary and secondary microphone signals are denoted $y_1(n)$ and $y_2(n)$ respectively. In respect with the explanations above and Figure 5.1 (b), we can write the following:

$$y_j(n) = s_j(n) + d_j(n) \tag{5.3}$$

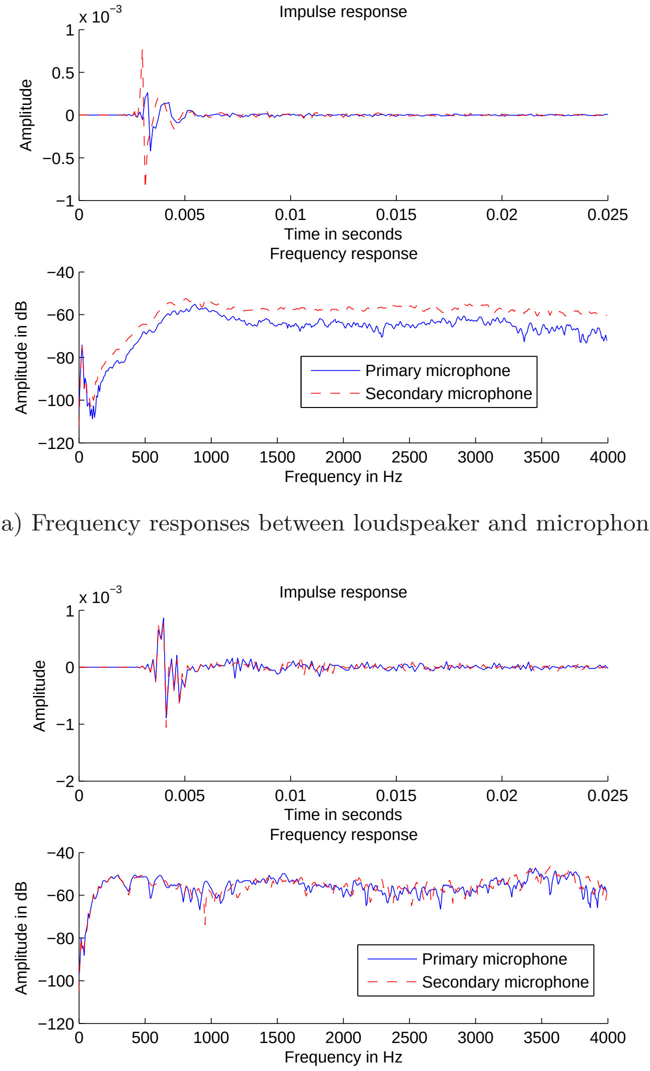
with $j \in \{1, 2\}$. In the following, the primary microphone refers the microphone which is placed further away from the loudspeaker i.e. with less power during echo-only periods.

The signal model presented here is quite general since it simply accounts for the physical interaction between the acoustic sources and transducers. The next step in our understanding of this model is to perform and analyze some recording. The recording setup includes both handset and handsfree scenarios.

5.1.2 Handsfree scenario analysis with mock-up phone

A mock-up phone has been built to get handsfree recordings. A detailed description of the mock-up phone used is furnished in Section 5.A. The mock-up phone consists of a solid plastic body equipped with a loudspeaker and two microphones. Microphones are placed in the bottom-top configuration (as in Figure 5.2 (b)). The mock-up phone is equipped with almost perfect transducers. This allows us to focus permits to focus the study on the acoustic interactions. Real mobile devices are mostly equipped with low-quality transducers. Since such transducers do not have flat frequency responses, it results that :

- the far-end signal emitted by the loudspeaker will be a modified version of the received signal
- the recorded microphone signals are modified versions of the signals picked by the microphones.



(a) Frequency responses between loudspeaker and microphones

(b) Frequency responses between artificial mouth and microphones

Figure 5.3: Frequency responses with the phone placed in front of the artificial mouth in a cabin environment

In addition, real mobile devices suffer from high non-linearities when used in handsfree. As result, the echo signal will be non-linear. Lastly, the electronic components somehow influence the acoustic interactions between the loudspeaker and the microphones. The use of the mock-up phone permits to reduce and eliminate all these undesirable effects and achieve a full assessment of the problem of linear echo.

The mock-up is used to measure impulse responses in different acoustic environments: cabin, office, meeting room (see Section 5.A for details about the recording setups). In all these experiments, the phone is placed such that the two microphones are approximately at equal distance from the artificial mouth. The recording setup in the cabin environment simulates scenarios in which the user holds the phone in his/her hands. For the office and meeting room environment, the device is placed on a table according to ITU-T Recommendation P.340 [ITU-T, 1996a], simulating a scenario in which the user places the device on a table to free his/her hand. For each of the acoustic environments considered, impulse responses $h_{1|2}(n)$ and $g_{1|2}(n)$ are measured through the exponential sine sweep technique [Farina, 2000]. An artificial head ([HEAD Acoustics HMS II.3]) with mouth simulator is used to simulate the near-end speaker and to accordingly to get $g_{1|2}(n)$.

Figure 5.3 (a) shows an example of frequency responses of the acoustic path between the loudspeaker and the microphones. This figure shows that the loudspeaker signal received by the microphones is not equally attenuated by the acoustic environment for each microphone. This implies that, during echo-only periods, the power of the signal on the secondary microphone is higher than that on the primary microphone. In [Habets, 2007], it is mentioned that the level of a sound wave at a given point is inversely proportional to the distance that separates this point from its source. The level difference observed in Figure 5.3 (a) is therefore in conformity with this property of sound wave propagation.

Figure 5.3 (b) shows an example of frequency responses between the artificial mouth and the microphones. We see that both impulse responses are very similar. These similarities can be explained by the position of the microphones compared to the artificial mouth.

5.1.3 Handset devices analysis with handset scenario

Mobile devices can also be used in handset. To complete our analysis of the DM echo problem, a real mobile device, the [Xolo X900], is used to record signals. The device used is equipped with one loudspeaker and two microphones which are still in bottom-top configuration. The far-end speech signal is played by the loudspeaker of the terminal and the near-end speaker is simulated by the same artificial head (HEAD Acoustics HMS II.3) as for the handsfree recording. The device is placed at the ear of the artificial head as described in the ITU-T Recommendation P.64 [ITU-T, 2007]. All recorded microphone signals contain echo-only, near-end only and double-talk (DT) periods.

In the analysis of the handset case, spectrograms and PSDs are used. An example of microphone signals is illustrated in Figure 5.4 and is composed of a near-end only period (from 0 to 9s) followed by an echo-only period (9s to the end). Based on these figures, we can state the following:

- During near-end only periods, the power of the signal on the primary microphone is higher than that on the secondary microphone:

$$\Phi^{y_2 y_2}(k,i) << \Phi^{y_1 y_1}(k,i) \tag{5.4}$$

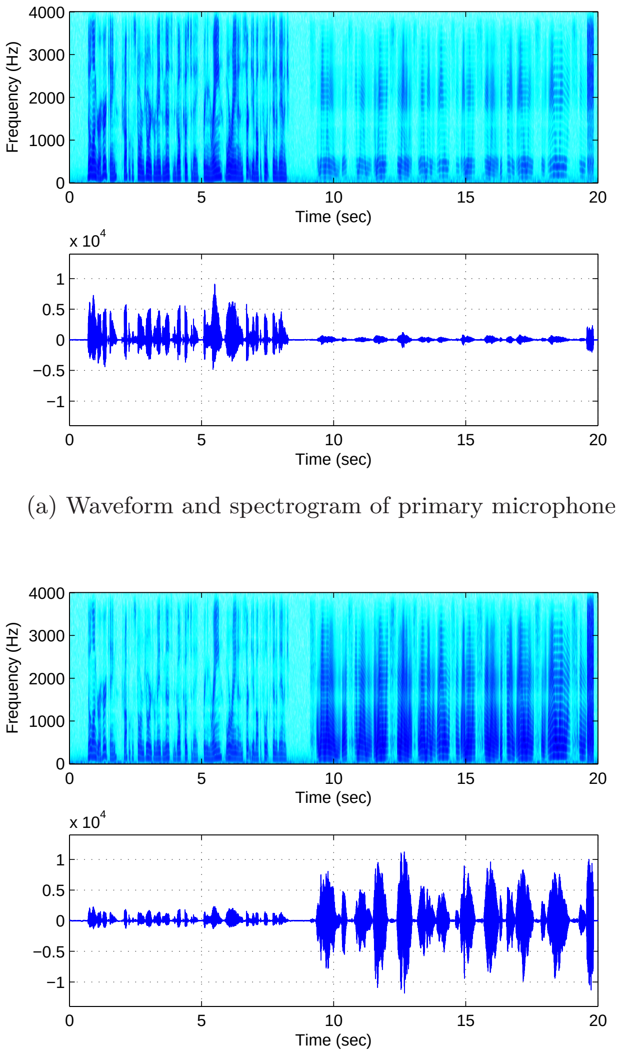
  where $\Phi^{y_j y_j}(k,i)$ is the PSD of the microphone signal $y_j(n)$ with $j \in \{1,2\}$.

- Following our formalism, the power of the signal on the primary microphone is lower than that on secondary microphone during echo-only periods:

$$\Phi^{y_1 y_1}(k,i) << \Phi^{y_2 y_2}(k,i). \tag{5.5}$$

  Once more the level difference observed is in agreement with sound wave propagation theory. Later on, we will exploit this power difference for the purpose of echo suppression

Data from the handset and handsfree cases are both used to assess proposed echo processing algorithms. Proposed algorithms are assessed in two steps: first with the



(a) Waveform and spectrogram of primary microphone

(b) Waveform and spectrogram of secondary microphone

Figure 5.4: Example of microphone signal in handset position

handsfree data and second with the handset data. Impulse responses measured are used to generate a database of test signals with different SERs. The database is used for extensive assessment of the proposed algorithms. The handset data will be used as a verification part for the proposed algorithms. This gives an insight on the performance of proposed algorithms on more realistic scenarios..

5.2 Proposed echo processing scheme

Our problem is to achieve echo cancellation given the two observations signals $y_j(n)$ (with $j \in \{1, 2\}$) and the reference signal $x(n)$. The system introduced here still outputs an estimate of the near-end speech signal $s_1(n)$ (i.e. near-end speech signal picked by the primary microphone) in an equivalent manner. The proposed echo cancellation schemes are composed of adaptive filtering followed by echo postfiltering and are illustrated in

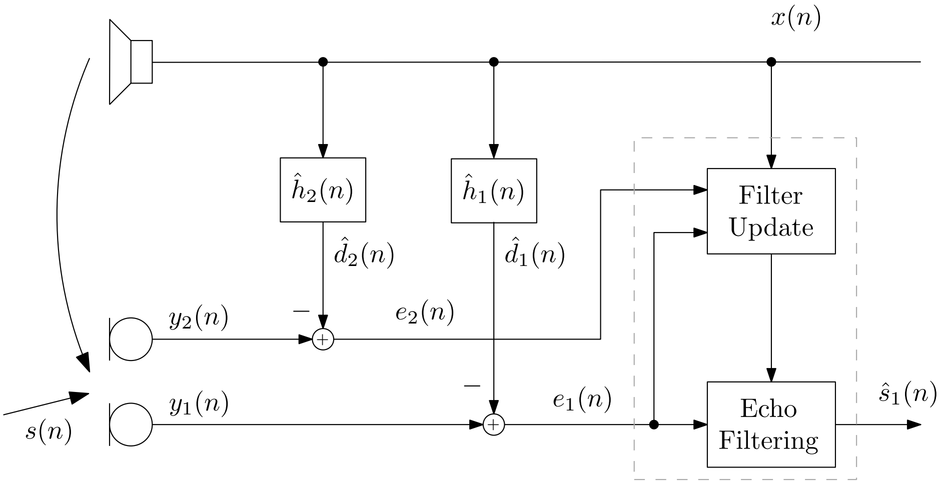


Figure 5.5: Proposed echo processing scheme

figures 5.5 and 5.6. The choice and position of each module is explained in the following.

5.2.1 Adaptive echo cancellation

In the system illustrated in Figure 5.5, adaptive echo cancellation (AEC) is composed of two adaptive filters: one adaptive filter per microphone path. The echo signal recorded by the each microphone is generated by the same loudspeaker. This means for each microphone signal, an estimate of the echo signal can be obtained through existing approaches to AEC. Similar use of the AEC for multi-microphone terminals can be found in the litterature [Guo et al., 2011, Jeannes et al., 2001, Kellermann, 1997]. AEC can be achieved through well-known existing approaches such as least mean square (LMS) or normalized LMS (NLMS) algorithms [Hänsler and Schmidt, 2004, Haykin, 2002]. Subband or frequency domain AEC algorithms can also be used.

The proposed DM echo processing scheme is designed such as to output an estimate of the near-end speech signal with an echo postfilter which is solely applied to one microphone path. AEC typically places high demand on memory and computational capacity. One could reduce the computational complexity of the system in Figure 5.5 by using one AEC instead of two as showed in Figure 5.6. The secondary microphone is directly input to the echo postfilter. In this way, the computational complexity is reduced however, the postfilter still exploits the dual-microphone architecture.

For the same reasons as in the SC case, some residual echo is present at the output of the AECs. The errors signals from the AEC can be expressed as follows:

$$e_j(n) = s_j(n) + \tilde{d}_j(n) \tag{5.6}$$

where $j \in \{1, 2\}$ and $\tilde{d}_j(n)$ represent the residual echo signal.

5.2.2 Echo postfiltering

As explained in Section 2.2.1.2, the postfilter is required to achieve further echo suppression. As schematized in Figures 5.5 and 5.6 and similarly as in SM terminals a frequency domain echo postfilter is used for residual echo suppression. Most frequency domain postfilters can be subdivided into two blocks: the filter update, in which the echo suppression

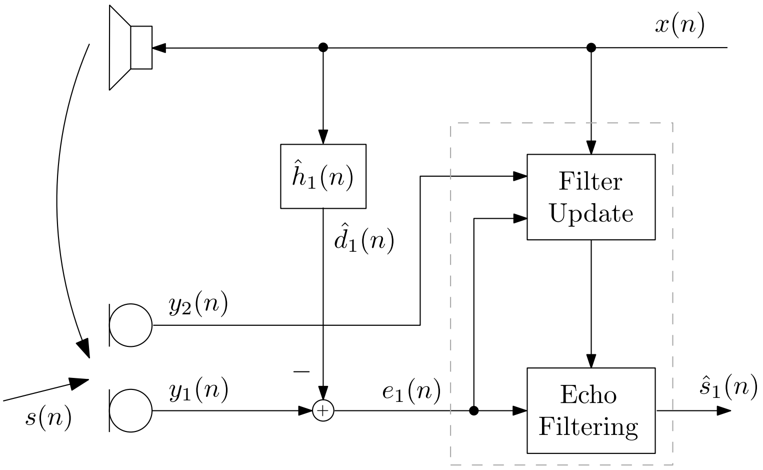


Figure 5.6: Alternative echo processing scheme with one AEC

filter is computed, and the echo suppression itself through filtering. The computation of the echo suppression filter use the input signals (loudspeaker and microphone signals) to compute an attenuation gain. This attenuation gain is then applied to the microphone path in the frequency domain or time domain to completely suppress the residual echo.

In our case, echo suppression is still applied only to the primary microphone path. This means existing echo suppression gain rules can still be used. Some example of gain rules which can be used in our case include [Hänsler and Schmidt, 2004, Haykin, 2002, **Yemdji** et al., 2010a]:

$$W_a(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)}, \quad W_b(k,i) = \frac{\Phi^{s_1 s_1}(k,i)}{\Phi^{s_1 s_1}(k,i) + \Phi^{\tilde{d}_1 \tilde{d}_1}(k,i)} \tag{5.7}$$

where $\Phi^{s_1 s_1}$ is the PSD of the near-end speech, $\Phi^{\tilde{d}_1 \tilde{d}_1}$ is the PSD of the residual echo at the primary microphone and ξ is the signal-to-echo ratio (SER) at the primary microphone. As explained in Chapter 2, both equations are mathematically equivalent but do not necessarily lead to the same results and speech quality.

In the postfilter used in the echo processing schemes in Figures 5.5 and 5.6, two microphone signals are used as inputs to the filter update module. Existing approaches to compute the echo suppression filter are based on single microphone. There is a need to design new estimation methods to compute the quantities involved in the computation of the echo postfilter gains. This can be done in two different ways:

- one simple option consists of keeping the existing gain rule computation unchanged (i.e. based on one microphone path) and simply use the microphones to add additional control on variables such as the gain values or echo PSD estimate. In this logic, the microphone signals can be used to design a double-talk detector(DTD). A DTD based on the level difference between the microphone signals is presented in Section 5.3.

- another option consists of using the dual microphone observations to design new estimations rules for the quantities involved in the computation of the echo suppression gain. In Section 5.4, we propose to use the level difference between the microphone signals to design new gain rule. Another approach to DM echo postfilter is presented in Chapter 6.

5.2.3 Synthesis

The target of the study of DM echo cancellation is to propose algorithms which improve echo cancellation performance while being implemented on real mobile devices. In the remainder of this thesis, only the echo processing in Figure 5.6 will be used because of its computational simplicity. Nevertheless, the proposed algorithms can be extended so to be used with two AECs. In the contributions presented in the remainder of this thesis, we consider that there is no ambient noise. The microphone signals only are only composed of the near-end speech and of the echo signals. This consideration permits a full in depth assessment of the behavior of proposed algorithms for echo processing.

5.3 Power level difference double-talk detector

In Section 5.1, we described the echo problem in dual-channel terminals and showed that important level differences can be observed depending on the active acoustic source (far-end speaker or near-end speaker). A very common tool to improve echo cancellation is the use of double-talk detection [Gänsler and Benesty, 2001][Huang et al., 2006, chap. 8]. In this section, we introduce a double talk detector (DTD) which exploits the level difference.

In practice, the level of a signal can be measured in terms of its amplitude, energy or power. In the following, the level is measured according to PSD but the formalism presented here can be applied to any other signal level measure. Since we are using the PSD, the level difference is referred to as the power level difference (PLD).

Section 5.3.1 presents our PLD based DTD while Section 5.3.2 carries on how this DTD can be used within an echo processing scheme.

5.3.1 Double-talk detector

In sections 5.1.2 and 5.1.3, we analyzed impulse responses for handsfree and microphone signals for handset (with bottom-top configuration as in Figure 5.2b) and showed that in both cases, important PLDs can be observed between the microphone signals in echo-only periods:

$$\Phi^{y_1 y_1}(k,i) << \Phi^{y_2 y_2}(k,i). \tag{5.8}$$

Assuming AEC does not amplify the echo, the PLD between the primary and secondary microphone paths is even more pronounced if it is measured after the adaptive filtering:

$$\Phi^{e_1 e_1}(k,i) \leq \Phi^{y_1 y_1}(k,i) \Rightarrow \Phi^{e_1 e_1}(k,i) << \Phi^{y_2 y_2}(k,i). \tag{5.9}$$

The assumption in Equation 5.9 is valid for most AECs in echo-only periods whether the AEC uses a fixed or variable stepsize. In DT periods, the convergence of the AEC depends a lot on the stepsize. Typically with fixed stepsize, the AEC will amplify the echo whereas with appropriate variable stepsize, the echo will not be amplified and might even be slightly attenuated. This means that in DT, typically

$$\Phi^{e_1 e_1}(k,i) \approx \Phi^{y_1 y_1}(k,i) \text{ or } \Phi^{e_1 e_1}(k,i) \geq \Phi^{y_1 y_1}(k,i) \Rightarrow \Phi^{e_1 e_1}(k,i) \geq \Phi^{y_2 y_2}(k,i). \tag{5.10}$$

Given this assumption about the AEC performance, we define the PLD as follows:

$$\Delta\Phi_{PLD}(k,i) = \Phi^{e_1 e_1}(k,i) - \Phi^{y_2 y_2}(k,i). \tag{5.11}$$

If we consider for example that only the far-end speaker is active, the PLD in Equation 5.11 can be rewritten as

$$\begin{aligned}\Delta\Phi_{PLD}(k,i) &= \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i) && (5.12)\\ &= \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) - \Phi^{d_2d_2}(k,i) && (5.13)\\ &= |\tilde{H}_1(k,i)|^2 \cdot \Phi^{xx}(k,i) - |H_2(k,i)|^2 \cdot \Phi^{xx}(k,i) && (5.14)\\ &= \left(|\tilde{H}_1(k,i)|^2 - |H_2(k,i)|^2\right) \cdot \Phi^{xx}(k,i). && (5.15)\end{aligned}$$

Let's consider two scenarios:

- First scenario: the loudspeaker signal has a certain level and its PSD is denoted $\Phi_0^{xx}(k,i)$. According to Equation 5.15, the PLD writes as:

$$\Delta\Phi_{PLD}(k,i) = \left(|\tilde{H}_1(k,i)|^2 - |H_2(k,i)|^2\right) \cdot \Phi_0^{xx}(k,i). \tag{5.16}$$

- Second scenario: the loudspeaker signal is an amplified version of that in the first scenario. Its PSD can be written as $b \cdot \Phi_0^{xx}(k,i)$ (with $b > 1$). In this case, the PLD writes as:

$$\Delta\Phi_{PLD}(k,i) = b \cdot \left(|\tilde{H}_1(k,i)|^2 - |H_2(k,i)|^2\right) \cdot \Phi_0^{xx}(k,i), \tag{5.17}$$

Equations 5.16 and 5.17 show that the PLD as defined in Equation 5.11 is dependent of the far-end speech level and of the echo return loss at each microphone which is not very convenient for DT detection. The echo return loss defines the energy loss between the loudspeaker signal and the echo signal at the microphone. It is typical of a device (direct path of the echo signal) and of the acoustic environment. To avoid being dependent of the voice level of the far-end and of the device, a normalized PLD is thus defined as follows:

$$\Delta\Phi_{PLD}(k,i) = \frac{\Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i)}{\Phi^{e_1e_1}(k,i) + \Phi^{y_2y_2}(k,i)}. \tag{5.18}$$

The normalization ensures that $\Delta\Phi_{PLD}(k,i)$ lies between -1 and 1 and is therefore more convenient to handle as we now have a measure which is independent of the voice level. In the remainder of this section, the behavior of this normalized PLD in echo-only, near-end-only and double-talk periods is studied.

5.3.1.1 During echo-only periods

During echo-only periods, the PSDs involved in the computation of the normalized PLD can be rewritten as functions of the loudspeaker signal PSD:

$$\begin{aligned}\Phi^{e_1e_1}(k,i) &= \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) = |\tilde{H}_1(k,i)|^2 \cdot \Phi^{xx}(k,i), && (5.19)\\ \Phi^{y_2y_2}(k,i) &= \Phi^{d_2d_2}(k,i) = |H_2(k,i)|^2 \cdot \Phi^{xx}(k,i). && (5.20)\end{aligned}$$

where $\tilde{H}_1(k,i) = H_1(k,i) - \hat{H}_1(k,i)$. From equations 5.19 and 5.20, the normalized PLD defined above can be rewritten as follows:

$$\Delta\Phi_{PLD}(k,i) = \frac{|\tilde{H}_1(k,i)|^2 - |H_2(k,i)|^2}{|\tilde{H}_1(k,i)|^2 + |H_2(k,i)|^2}. \tag{5.21}$$

By defining the echo relative transfer function (RTF) as follows:

$$\Gamma(k,i) = \frac{H_2(k,i)}{\tilde{H}_1(k,i)}, \tag{5.22}$$

Equation 5.21 can be rewritten as:

$$\Delta\Phi_{PLD}(k,i) = \frac{1 - |\Gamma(k,i)|^2}{1 + |\Gamma(k,i)|^2}. \tag{5.23}$$

In Section 5.1, we showed that in echo-only periods, there is an important PLD between the primary and secondary microphone signals. Assuming as explained above that the AEC does not amplify echo, it can be stated that:

$$\left|\tilde{H}_1(k,i)\right| << \left|H_2(k,i)\right| \Longrightarrow \left|\Gamma(k,i)\right| \to +\infty. \tag{5.24}$$

From equations 5.23 and 5.24, it can be deduced that the normalized PLD tends to -1 during echo-only periods. It is possible to go further in our analysis by stating that the larger $\left|\Gamma(k,i)\right|$, the closer the normalized PLD value will be to -1. Given the definition of the echo RTF, two possible methods can be used to guarantee high values of the echo RTF $\left|\Gamma(k,i)\right|$:

- Assure that $\left|\tilde{H}_1(k,i)\right|$ is as close as possible to zero. This supposes that the AEC should converge very quickly and that the residual echo at its output should be very small. In Chapter 2, we explain why the implementation of such an AEC is not feasible in practice.

- Place the secondary microphone such that $\left|H_2(k,i)\right|$ is as high as possible. Such behavior can be obtained by placing the secondary microphone as closed as possible to the loudspeaker. Later on, in our experiments, we analyze the impact of the position of the secondary microphone.

5.3.1.2 During near-end only periods

During near-end-only periods, the PSDs involved in the computation of the normalized PLD are functions of the near-end speech signal $s(n)$:

$$\Phi^{e_1e_1}(k,i) = \Phi^{s_1s_1}(k,i) = |G_1(k,i)|^2 \cdot \Phi^{ss}(k,i), \tag{5.25}$$

$$\Phi^{y_2y_2}(k,i) = \Phi^{s_2s_2}(k,i) = |G_2(k,i)|^2 \cdot \Phi^{ss}(k,i). \tag{5.26}$$

Similarly as for the echo-only case, the near-end RTF can be defined

$$\Theta(k,i) = \frac{G_2(k,i)}{G_1(k,i)}, \tag{5.27}$$

and used to rewrite the normalized PLD as follows:

$$\Delta\Phi_{PLD}(k,i) = \frac{1 - |\Theta(k,i)|^2}{1 + |\Theta(k,i)|^2}. \tag{5.28}$$

Referring to the signal recording presented above, two cases can be considered:

- In handset typically,

$$|G_2(k,i)| << |G_1(k,i)| \Longrightarrow |\Theta(k,i)| \to 0 \tag{5.29}$$

  which leads to values of PLD close to +1.

- In handsfree mode, the behavior is slightly different as

$$|G_2(k,i)| \approx |G_1(k,i)| \Longrightarrow |\Theta(k,i)| \to 1 \tag{5.30}$$

  therefore leads to values of PLD close to 0.

5.3.1.3 Double-talk

Using the RTFs defined above, the PSD of the secondary microphone can be rewritten as follows:

$$\Phi^{y_2y_2}(k,i) = \Phi^{s_2s_2}(k,i) + \Phi^{d_2d_2}(k,i) \tag{5.31}$$
$$\Phi^{y_2y_2}(k,i) = |G_2(k,i)|^2 \cdot \Phi^{ss}(k,i) + |H_2(k,i)|^2 \cdot \Phi^{xx}(k,i) \tag{5.32}$$
$$\Phi^{y_2y_2}(k,i) = |\Theta(k,i)|^2 \cdot \Phi^{s_1s_1}(k,i) + |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \tag{5.33}$$

By introducing Equation 5.33 in that of the PLD in Equation 5.18 and considering DT periods, one obtains:

$$\Delta\Phi_{PLD}(k,i) = \frac{(1-|\Gamma(k,i)|^2)\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + (1-|\Theta(k,i)|^2)\Phi^{s_1s_1}(k,i)}{(1+|\Gamma(k,i)|^2)\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + (1+|\Theta(k,i)|^2)\Phi^{s_1s_1}(k,i)} \tag{5.34}$$
$$\Delta\Phi_{PLD}(k,i) = \frac{(1-|\Gamma(k,i)|^2) + (1-|\Theta(k,i)|^2)\cdot\xi(k,i)}{(1+|\Gamma(k,i)|^2) + (1+|\Theta(k,i)|^2)\cdot\xi(k,i)} \quad \text{with} \quad \xi(k,i) = \frac{\Phi^{s_1s_1}(k,i)}{\Phi^{\tilde{d}_1\tilde{d}_1}(k,i)}. \tag{5.35}$$

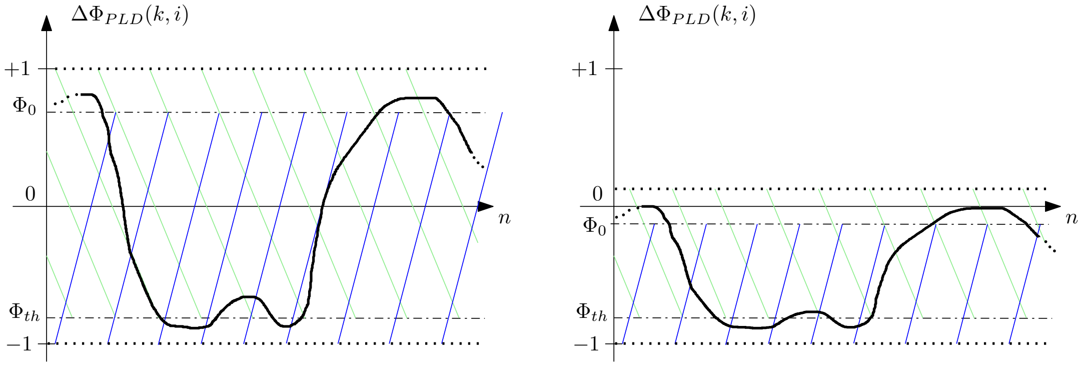
From Equation 5.35 and the signal recording presented in Section 5.1, the following conclusions can be drawn about the PLD values during double-talk:

- **Handset mode:** Typically, $\Theta(k,i) \to 0$ while $\Gamma(k,i) \to +\infty$. By simplification, the PLD can be approximated as:

$$\Delta\Phi_{PLD}(k,i) \approx \frac{-|\Gamma(k,i)|^2 + \xi(k,i)}{|\Gamma(k,i)|^2 + \xi(k,i)}. \tag{5.36}$$

  Depending on the $\xi(k,i)$ values, the PLD values will be positive or negative but still be bounded between -1 and +1. The approximation in Equation 5.36 also shows that:

  - as the $\xi(k,i)$ tends to zero (i.e when residual echo is dominant compared to near-end speech), the PLD values will be closer to -1 as for the echo-only periods. This is not very problematic as we are effectively in echo-only given the fact that $\xi(k,i)$ tends to zero.
  - In contrast, as the SER $\xi(k,i)$ increases (i.e when the residual echo level becomes comparable to that of the near-end), the further PLD values will be from -1.



(a) Typical PLD for handset scenario

(b) Typical PLD for handsfree scenario

Figure 5.7: Illustration of the PLD behavior

- **Handsfree mode:** The behavior is slightly different as $\Theta(k,i) \to 1$ while $\Gamma(k,i) \to +\infty$ and by simplification the PLD can be approximated as follows:

$$\Delta\Phi_{PLD}(k,i) \approx \frac{-|\Gamma(k,i)|^2}{|\Gamma(k,i)|^2 + 2 \cdot \xi(k,i)}. \tag{5.37}$$

  From Equation 5.37, we can see that the PLD values will mostly be comprised between -1 and 0. This approximation also emphasizes the fact that depending on the values of $\xi(k,i)$, the PLD values will be more or less closer to 0.

Equations 5.36 and 5.37 show that PLD values can be used to distinguish echo-only from double-talk periods.

5.3.1.4 Synthesis on the behavior of normalized PLD

The analysis of the a-priori behavior of the PLD for handset and handsfree use-cases is synthesized in Figure 5.7. For handsfree scenarios, PLD values are comprised in between -1 and 0 whereas for handset scenarios, PLD values are comprised between -1 and +1. For both handset and handsfree modes, PLD values are close to -1 during echo-only periods.

The aim of the normalized PLD defined in Equation 5.18 is to differentiate echo-only periods from double-talk periods. Although the PLD is analyzed for near-end only periods in Section 5.3.1.2, detecting near-end only periods is not really a problem in real-time systems. Near-end only periods are periods during which the microphone signal has speech while the loudspeaker signal has no speech. In a real time system, near-end only periods can be detected by applying a voice activity detector to the microphone signal on one hand and to the loudspeaker signal on the other hand. From Figure 5.7, we can conclude that echo-only periods can be detected by applying a threshold to the PLD values:

$$\text{Echo-only periods if} \quad \Delta\Phi_{PLD}(k,i) < \Phi_{th}(i) \quad \text{with} \quad \Phi_{th}(i) > -1 \tag{5.38}$$

where $\Phi_{th}(i)$ is the threshold which can be frequency-depended or the same for all frequency bins. Double-talk periods are then detected by combining a far-end speech activity detector with the above defined PLD:

$$\Delta\Phi_{PLD}(k,i) \geq \Phi_{th}(i) \quad \text{AND} \quad \text{far end speech active}. \tag{5.39}$$

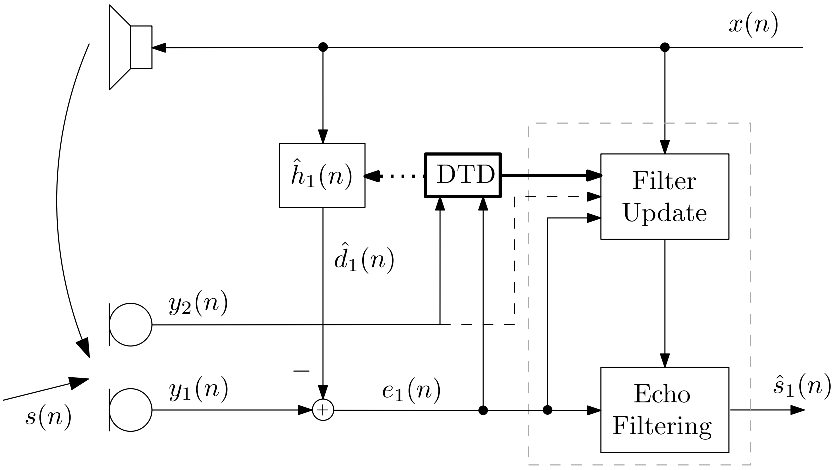


Figure 5.8: DM echo cancellation scheme including PLD based DTD control

A very simple far-end speech activity detection consists of thresholding the far-end PSD:

$$\Phi^{xx}(k,i) > \Phi_0^{xx} \tag{5.40}$$

where Φ_0^{xx} is a threshold above which speech is assumed to be present. In our case, we assume there is no noise in the acoustic environment: this threshold is set to be fixed. In presence of noise, some techniques exits to account for the noise level in the this threshold [Davis et al., 2006, Tanyer and Ozer, 2000].

The frequency-domain based DTD decision making is of interest if we imagine that the DTD is done after a subband AEC or that we have an a-priori knowledge on the echo return loss of the device. Subband AEC might typically result in echo suppression performance that differs from one subband to another. The threshold might then be set according the AEC performance in each subband. In the same way, a-priori knowledge on the echo return loss of our system might inform us on the shape of the frequency response of the echo path. The value of threshold might then be set according to the echo return loss. Such a-priori information can be of interest when tuning the DTD for a given device.

In our implementation of the DTD, although the normalized PLD is computed in the frequency domain, the DTD threshold $\Phi_{th}(i)$ has the same value for the whole range of frequencies. Even by setting $\Phi_{th}(i)$ independently of the frequency bin, our implementation still gain advantage of the fact that the normalized is computed in the frequency domain. If we consider a case where the far- and near-end speakers do not have the same pitch frequencies, DT will only occur for some frequencies (i.e. frequencies that are multiple of both pitch frequencies). Our frequency domain DTD will still be able to detect specific frequencies where double-talk occurs. In comparison to fullband DTDs, our DTD will output finer DT decision. In the next section, we present how our DTD is used to monitor the echo processing.

5.3.2 Usage within proposed echo processing scheme

Figure 5.8 illustrates how the PLD based DTD presented above can be used within an echo cancellation scheme. We see that the DTD output is fed into the AEC and the echo postfilter. In contrast to the scheme illustrated in Figure 5.6, the postfilter does not necessarily use the DM signals. The proposed DTD is based on DM signals but can be used with a single microphone postfilter. This means our DTD can be used to improve any SM echo postfilter. Later on we show how the proposed DTD can be used to improve the performance of the echo postfilter used in Chapter 3.

A typical use of the proposed DTD within AEC is to freeze the adaptation when double-talk is detected, i.e. setting the step-size (which can be variable or fixed) to 0. In some cases, like in Chapter 4, the AEC might not take place in the same frequency or subband domain as the proposed DTD. Assuming the AEC operates in the fullband domain, a fullband DTD can be obtained by averaging the PLD values along frequencies. In this way for a time instance k, we obtain one value for the PLD which we use for the DTD. Alternatively, a DTD for a frequency band of interest can be derived from the normalized PLD by averaging the PLD values of a set of frequency bins or by alternatively computing the PLD using energies (instead of the PSDs as done until now).

As illustrated in Figure 5.8, the DTD can be used to control echo postfiltering. One method could be to postprocess the echo suppression gain (after it has been updated according to the formalism of our choice) by setting it to its minimum value during echo-only periods

$$\text{if } \Delta\Phi_{PLD}(k,i) < \Phi_{th}(i) \Longrightarrow W(k,i) = W_{min} \tag{5.41}$$

where W_{min} is the echo suppression floor value. This permits to achieve maximum echo suppression during echo-only periods. We can also imagine a more indirect modification/improvement of the echo postfilter by influencing variables such as the echo PSD required in the computation. During echo-only periods, the residual echo PSD $\Phi^{\tilde{d}_1\tilde{d}_1}(k,i)$ is set to be equal to the error signal PSD.

$$\text{if } \Delta\Phi_{PLD}(k,i) < \Phi_{th}(i) \Longrightarrow \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) = \Phi^{e_1e_1}(k,i) \tag{5.42}$$

The post-processing in Equation 5.42 permits to improve echo suppression by improving the accuracy of the echo PSD estimate. Parameters such as the overestimation factor involved in the computation of echo PSD estimate can also be controlled using the PLD values. As explained in Chapter 2, it is very common to weight the echo PSD estimate with an overestimation factor which aims to compensate for estimation errors. High values of overestimation lead to high echo suppression and high near-end speech distortions while small values lead to the opposite effects. With our frequency domain based DTD, we can introduce a two-step overestimation factor η between which we would switch depending on whether we are in echo-only or double talk:

$$\begin{aligned} \text{if } \Delta\Phi_{PLD}(k,i) < \Phi_{th}(i) \Longrightarrow \eta &= \eta_{echo} \\ \text{otherwise } \eta &= \eta_{dt} \end{aligned} \tag{5.43}$$

with $\eta_{echo} > \eta_{dt}$.

5.4 Power level difference based echo suppression gain rule

In Section 5.2, we mentioned that the DM signals could be used for echo postfiltering. In this section, we introduce a novel echo suppression gain rule. The proposed gain rule is derived from the Wiener gain and uses both the error signal from the AEC $e_1(n)$ and the secondary microphone signal $y_2(n)$. The proposed gain rule does not require an explicit estimate of the residual echo PSD but is based on relative transfer functions (RTFs).

This section is organized as follows. In Section 5.4.1, we introduce our gain rule. Section 5.4.2 deals about approaches to estimate the RTFs required in the computation of our gain rule. In Section 5.4.3, we analyze the behavior of our gain by studying the impact of the error in the RTF estimates. Lastly, Section 5.4.4 summarizes the computation of the gain rule.

5.4.1 Gain rule

The proposed gain rule is derived from the Wiener echo suppression gain rule which expresses as follows:

$$W(k,i) = \frac{\Phi^{s_1 s_1}(k,i)}{\Phi^{s_1 s_1}(k,i) + \Phi^{\tilde{d}_1 \tilde{d}_1}(k,i)}. \tag{5.44}$$

The computation of the gain specified in Equation 5.44 requires the estimation of the near-end and residual echo PSDs $\Phi^{s_1 s_1}$ and $\Phi^{\tilde{d}_1 \tilde{d}_1}$. An alternative computational method for the gain rule in Equation 5.44 is presented. The method proposed exploits the PLD between the microphone signals and results in an implementation that does not require the estimation of $\Phi^{s_1 s_1}$ and $\Phi^{\tilde{d}_1 \tilde{d}_1}$.

Assuming the loudspeaker and near-end speech signal are independent (i.e. $\Phi^{xs} = 0$), the PSD of the microphone signals that input the postfilter are defined as follows:

$$\Phi^{e_1 e_1}(k,i) = \Phi^{s_1 s_1}(k,i) + \Phi^{\tilde{d}_1 \tilde{d}_1}(k,i) \tag{5.45}$$

$$\Phi^{y_2 y_2}(k,i) = \Phi^{s_2 s_2}(k,i) + \Phi^{d_2 d_2}(k,i). \tag{5.46}$$

Equations 5.45 and 5.46 can equivalently be written as functions of the loudspeaker and near-end speech PSDs Φ_{xx}, Φ_{ss}:

$$\Phi^{e_1 e_1}(k,i) = |G_1(k,i)|^2 \cdot \Phi^{ss}(k,i) + |\tilde{H}_1(k,i)|^2 \cdot \Phi^{xx}(k,i) \tag{5.47}$$

$$\Phi^{y_2 y_2}(k,i) = |G_2(k,i)|^2 \cdot \Phi^{ss}(k,i) + |H_2(k,i)|^2 \cdot \Phi^{xx}(k,i). \tag{5.48}$$

Using the echo and near-end RTFs Γ and Θ defined in Equations 5.22 and 5.27, the secondary microphone PSD can be rewritten as

$$\Phi^{y_2 y_2}(k,i) = |\Theta(k,i)|^2 \cdot \Phi^{s_1 s_1}(k,i) + |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1 \tilde{d}_1}(k,i). \tag{5.49}$$

Let's define two new PLDs:

$$\Delta\Phi_{PLD}^{echo}(k,i) = |\Theta(k,i)|^2 \cdot \Phi^{e_1 e_1}(k,i) - \Phi^{y_2 y_2}(k,i) \tag{5.50}$$

$$\Delta\Phi_{PLD}^{near}(k,i) = \Phi^{y_2 y_2}(k,i) - |\Gamma(k,i)|^2 \cdot \Phi^{e_1 e_1}(k,i). \tag{5.51}$$

Using Equations 5.45 and 5.49, the PLDs in Equations 5.50 and 5.51 can be rewritten as follows

$$\Delta\Phi_{PLD}^{echo}(k,i) = \left(|\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1 \tilde{d}_1}(k,i) \tag{5.52}$$

$$\Delta\Phi_{PLD}^{near}(k,i) = \left(|\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{s_1 s_1}(k,i). \tag{5.53}$$

Equations 5.52 and 5.53 show that the PLDs in equations 5.50 and 5.51 are respectively functions of the residual echo and near-end speech PSDs present in the primary microphone path. Using these PLDs, the Wiener gain rule in 5.44 can be redefined as follows:

$$\hat{W}(k,i) = \frac{\Phi^{s_1 s_1}(k,i)}{\Phi^{s_1 s_1}(k,i) + \Phi^{\tilde{d}_1 \tilde{d}_1}(k,i)} = \frac{\Delta\Phi_{PLD}^{near}(k,i)}{\Delta\Phi_{PLD}^{near}(k,i) + \Delta\Phi_{PLD}^{echo}(k,i)}. \tag{5.54}$$

One can refer to **Yemdji** et al. [2013] for an extension of the proposed DM echo postfilter gain rule to multimicrophone architectures. In a real time implementation, the computation of the proposed echo suppression gain rule according to 5.54 only requires the estimation of the relative transfer functions (RTFs) Γ and Θ which we discussed next.

5.4.2 Relative transfer function estimation

5.4.2.1 Near-end relative transfer function estimation

The near-end RTF Θ is defined as the ratio of the frequency responses of the acoustic paths between the near-end speaker and each microphone:

$$\Theta(k,i) = \frac{G_2(k,i)}{G_1(k,i)}. \tag{5.55}$$

Θ can also be interpreted as a gain such that

$$s_2(k,i) = \Theta(k,i) \cdot s_1(k,i). \tag{5.56}$$

Considering near-end only speech activity period (i.e. $e_1(k,i) = s_1(k,i) = G_1(k,i) \cdot s(k,i)$ and $y_2(k,i) = s_2(k,i) = G_2(k,i) \cdot s(k,i)$), an estimate $\hat{\Theta}$ of Θ can be obtained through mean square error (MSE) or least square error (LSE) minimization. The minimum MSE (MMSE) criteria used for the derivation of the MMSE estimate of $\hat{\Theta}$ is expressed as follows:

$$\hat{\Theta}(k,i) = \underset{\Theta}{\operatorname{argmin}} \left|S_2(k,i) - \hat{S}_2(k,i)\right| \quad \text{with} \quad \hat{S}_2(k,i) = \hat{\Theta}(k,i) \cdot S_1(k,i) \tag{5.57}$$

and leads to the following

$$\hat{\Theta}_{MMSE}(k,i) = \frac{\Phi^{e_1 y_2}(k,i)}{\Phi^{y_2 y_2}(k,i)}. \tag{5.58}$$

In [Shalvi and Weinstein, 1996], the above MMSE estimate is shown to be biased and an alternative unbiased estimator is proposed. This estimator exploits speech non-stationarity through least square method and leads to the following:

$$\hat{\Theta}_{LS} = \frac{\left\langle \Phi^{e_1 e_1} \cdot \Phi^{e_1 y_2} \right\rangle - \left\langle \Phi^{e_1 e_1} \right\rangle \cdot \left\langle \Phi^{e_1 y_2} \right\rangle}{\left\langle (\Phi^{e_1 e_1})^2 \right\rangle - \left\langle \Phi^{e_1 e_1} \right\rangle^2} \tag{5.59}$$

where $\left\langle \Phi^{y_j y_l} \right\rangle = \frac{1}{K} \sum_{k=1}^{K} \Phi^{y_j y_l}(k,i), (j,l) \in \{1,2\}$ is an averaging operator over time and K is the averaging window length which we set to 13 as in [Gannot et al., 2001]. The near-end RTF estimate in Equation 5.59 requires an important computational capacity compared to that in Equation 5.59. Part of the experiments reported later in this chapter present a comparative assessment of the impact of these two estimates.

In our implementation, both near-end RTF estimates are updated during near-end speech only activity periods. Intervals containing only near-end speech are detected using two thresholds:

- one for the loudspeaker signal energy
- another for the energies of the two microphones.

The threshold on the loudspeaker energy permits to avoid adaptation during far-end activity periods whereas the threshold on the microphone signals permits to avoid adaptation during near-end silence periods or using low amplitude microphone signals.

5.4.2.2 Echo relative transfer function estimation

The echo RTF Γ is defined as the ratio between the secondary and the primary residual echo paths

$$\Gamma(k,i) = \frac{H_2(k,i)}{\tilde{H}_1(k,i)}. \tag{5.60}$$

Similarly as for the near-end RTF Θ in Equation 5.56 , Γ can be seen as a function that defines the link between the residual echo of primary and secondary microphone in the following manner

$$d_2(k,i) = \Gamma(k,i) \cdot \tilde{d}_1(k,i). \tag{5.61}$$

Using the above equation, the error and microphone signals e_1 and y_2 can be rewritten as:

$$e_1(k,i) = s_1(k,i) + \tilde{d}_1(k,i) \qquad = G_1(k,i) \cdot s(k,i) + \tilde{H}_1(k,i) \cdot x(k,i) \tag{5.62}$$

$$y_2(k,i) = s_2(k,i) + \Gamma(k,i) \cdot \tilde{d}_1(k,i) \qquad = G_2(k,i) \cdot s(k,i) + H_2(k,i) \cdot x(k,i). \tag{5.63}$$

Using the fact that $\tilde{d}_1$ and d_2 are both generated by the loudspeaker and the independence between the loudspeaker and near-end speech signal(i.e. $\Phi^{xs} = 0$), Γ can be estimated through the cross-correlation between the loudspeaker and the microphones signals:

$$\Phi^{xe_1}(k,i) = \tilde{H}_1(k,i) \cdot \Phi^{xx}(k,i) \quad \text{and} \quad \Phi^{xy_2}(k,i) = H_2(k,i) \cdot \Phi^{xx}(k,i) \tag{5.64}$$

$$\hat{\Gamma}(k,i) = \frac{\Phi^{xy_2}(k,i)}{\Phi^{xe_1}(k,i)}. \tag{5.65}$$

5.4.3 Analysis of the proposed gain rule

The proposed gain rule is directly dependent on the echo and near-end RTF estimates. Errors in the RTFs estimate will disturb the gain rule behavior. Prior to assessing the performance of the proposed gain rule in terms of echo suppression, we analyze how these errors impact the new gain rule. For this analysis, the echo and near-end RTF estimates are modeled as the sum of their real value with and of additive noise which accounts for estimation errors:

$$\hat{\Gamma}(k,i) = \Gamma(k,i) + \Delta\Gamma(k,i) \tag{5.66}$$

$$\hat{\Theta}(k,i) = \Theta(k,i) + \Delta\Theta(k,i) \tag{5.67}$$

where $\Delta\Theta$ and $\Delta\Gamma$ represent the errors in the echo and near-end RTF estimates respectively. Echo suppression takes place during echo-only and double-talk periods which we considered in our analysis.

Errors in the echo and near-end RTFs estimates are analyzed separately in sections 5.4.3.1 and 5.4.3.2 respectively. The behavior of the proposed gain is them summarized in 5.4.3.3.

5.4.3.1 Impact of the echo RTF

In this section, we analyze the impact of the echo RTF on the proposed gain rule. For this, we rewrite its estimate $\hat{\Gamma}$ as in Equation 5.66. Given this definition of the echo RTF, the PLDs used in the computation of the new gain rule become

$$\Delta\Phi^{echo}_{PLD}(k,i) = |\Theta(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i) \tag{5.68}$$

$$\begin{aligned} \Delta\Phi^{near}_{PLD}(k,i) &= \Phi^{y_2y_2}(k,i) - |\hat{\Gamma}(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) \\ &= \Phi^{y_2y_2}(k,i) - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i). \end{aligned} \tag{5.69}$$

In echo-only periods, the PSDs $\Phi^{e_1e_1}$ and $\Phi^{y_2y_2}$ are written as:

$$\Phi^{e_1e_1}(k,i) = \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) \tag{5.70}$$

$$\Phi^{y_2y_2}(k,i) = \Phi^{d_2d_2}(k,i) = |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \tag{5.71}$$

By rewriting the PLDs in Equations 5.68 and 5.69 with the PSDs in Equations 5.70 and 5.71, we obtain the following:

$$\begin{aligned}\Delta\Phi^{echo}_{PLD}(k,i) &= |\Theta(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) - |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\\ &= \left(|\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\end{aligned} \tag{5.72}$$

$$\begin{aligned}\Delta\Phi^{near}_{PLD}(k,i) &= |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\\ &= \left(|\Gamma(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i).\end{aligned} \tag{5.73}$$

Using equations 5.72 and 5.73, the proposed echo suppression gain therefore becomes

$$W(k,i) = \frac{\Delta\Phi^{near}_{PLD}(k,i)}{\Delta\Phi^{near}_{PLD}(k,i) + \Delta\Phi^{echo}_{PLD}(k,i)} \tag{5.74}$$

$$= \frac{\left(|\Gamma(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)}{\left(|\Gamma(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \left(|\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)}$$

The PSD $\Phi^{\tilde{d}_1\tilde{d}_1}(k,i)$ in the numerator and denominator can be simplified and the echo suppression gain $W(k,i)$ may be simplified to:

$$W(k,i) = \frac{|\Gamma(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2}{|\Gamma(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2 + \left(|\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right)} \tag{5.75}$$

$$= \frac{|\Gamma(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2}{|\Theta(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2} = W_0(k,i) \tag{5.76}$$

In the ideal case during echo-only periods, the echo suppression gains should be equal to zero (i.e. $W(k,i) = 0$) to completely suppress the residual echo. Instead, we see that the residual echo will only be attenuated and might still be audible after postfiltering. Nevertheless, it is of interest to note that, the smaller the estimation error $\Delta\Gamma$, the more the numerator of W_0 tends to zero thus resulting in better echo suppression. One can also note that W_0 is independent of both the echo and the near-end speech. Instead W_0 is only dependent on the RTFs and more specifically in the error on the echo RTF. To some extend, W_0 can be seen as a spectral floor. Let's consider the computation of the echo RTF as presented in Section 5.4.2.2. When the required cross-PSDs are computed through autoregressive smoothing, they need some time to reach their steady state. The error on the echo RTF will be important during the transient period and relatively low during the steady state period. Typical at the beginning of an echo-only period, $\Delta\Gamma$ will be high and by so doing W_0 will be high and will not achieve enough echo attenuation. When the echo RTF estimate has reached its steady state, however, although $\Delta\Gamma$ will be quite low, W_0 will be higher zero and will in fact result in flooring of the amount of echo suppression.

Equations 5.74 to 5.76 show that the errors that occur in the echo RTF estimate in echo-only periods results in a spectral floor of echo attenuation. We can already state that the amount of echo suppression during echo-only periods will therefore be less than that expected because of this spectral floor.

Now considering double-talk periods, the PLDs required in the computation of the echo suppression gains are expressed as follows

$$\begin{aligned}\Delta\Phi_{PLD}^{echo}(k,i) =&|\Theta(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i) && (5.77)\\ =&|\Theta(k,i)|^2 \cdot \left(\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \Phi^{s_1s_1}(k,i)\right)\\ &- \left(|\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + |\Theta(k,i)|^2 \cdot \Phi^{s_1s_1}(k,i)\right)\\ =&\left(|\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \left(|\hat{\Theta}(k,i)|^2 - |\Theta(k,i)|^2\right) \cdot \Phi^{s_1s_1}(k,i)\\ =&\left(|\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) && (5.78)\end{aligned}$$

$$\begin{aligned}\Delta\Phi_{PLD}^{near}(k,i) =& \Phi^{y_2y_2}(k,i) - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) && (5.79)\\ =&\left(|\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + |\Theta(k,i)|^2 \cdot \Phi^{s_1s_1}(k,i)\right)\\ &- |\Gamma(k,i) + \Delta\Gamma(k,i)|^2 \cdot \left(\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \Phi^{s_1s_1}(k,i)\right)\\ =&\left(|\Gamma(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\\ &+ \left(|\hat{\Theta}(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2\right) \cdot \Phi^{s_1s_1}(k,i). && (5.80)\end{aligned}$$

Using equations 5.78 and 5.80, the echo suppression gain can be rewritten as:

$$\hat{W}(k,i) = \frac{\Delta\Phi_{PLD}^{near}(k,i)}{\Delta\Phi_{PLD}^{near}(k,i) + \Delta\Phi_{PLD}^{echo}(k,i)} = W(k,i) + \Delta W(k,i) \tag{5.81}$$

where the additional term ΔW is expressed as follows:

$$\begin{aligned}\Delta W(k,i) =& \frac{\left(|\Gamma(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)}{\left(|\Theta(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2\right) \cdot \left(\Phi^{s_1s_1}(k,i) + \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\right)} && (5.82)\\ =& W_0(k,i) \cdot \frac{\Phi^{\tilde{d}_1\tilde{d}_1}(k,i)}{\Phi^{s_1s_1}(k,i) + \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)} && (5.83)\\ =& W_0(k,i) \cdot \frac{1}{1+\xi(k,i)} \quad \text{with} \quad \xi(k,i) = \frac{\Phi^{s_1s_1}(k,i)}{\Phi^{\tilde{d}_1\tilde{d}_1}(k,i)}. && (5.84)\end{aligned}$$

The error ΔW is always positive: the derivation of W_0 in Equations 5.74 to 5.76 shows $W_0 > 0$ and the PSD terms are always positive. The echo suppression gain is expressed as the sum of its real value to which is added an estimation error ΔW. The error term ΔW is inversely proportional to the near-end to residual echo ratio ξ:

- For high values of ξ, ΔW will tend to zero

$$\xi(k,i) \to +\infty \Longrightarrow \Delta W(k,i) \to 0 \quad \text{and} \quad \hat{W}(k,i) \approx W(k,i). \tag{5.85}$$

- For small values of ξ as it is the case for double-talk periods, ΔW will tend to W_0

$$\xi(k,i) \to 0 \Longrightarrow \Delta W(k,i) \to W_0(k,i) \quad \text{and} \quad \hat{W}(k,i) \approx W(k,i) + W_0(k,i). \tag{5.86}$$

We see clearly from Equation 5.86, the proposed gain rule will result in less echo attenuation than that would be expected during double-talk periods.

5.4.3.2 Impact of the near-end RTF

Similarly as in Section 5.4.3.1, we assume the near-end RTF estimates as follows:

$$\hat{\Theta}(k,i) = \Theta(k,i) + \Delta\Theta(k,i) \tag{5.87}$$

where $\Delta\Theta$ accounts for estimation errors. Considering echo-only periods, the PLDs required in the computation of the echo suppression gain are written as follows:

$$\begin{aligned}
\Delta\Phi_{PLD}^{echo}(k,i) &= |\hat{\Theta}(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i) && (5.88)\\
&= |\Theta(k,i) + \Delta\Theta(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) - |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\\
&= \left(|\Theta(k,i) + \Delta\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) && (5.89)
\end{aligned}$$

$$\begin{aligned}
\Delta\Phi_{PLD}^{near}(k,i) &= \Phi^{y_2y_2}(k,i) - |\Gamma(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) && (5.90)\\
&= |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) - |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\\
&= 0. && (5.91)
\end{aligned}$$

Using equations 5.89 and 5.91, the proposed gain rule becomes:

$$\begin{aligned}
W(k,i) &= \frac{\Delta\Phi_{PLD}^{near}(k,i)}{\Delta\Phi_{PLD}^{near}(k,i) + \Delta\Phi_{PLD}^{echo}(k,i)} && (5.92)\\
&= \frac{0}{0 + \left(|\Theta(k,i) + \Delta\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)} = 0 && (5.93)
\end{aligned}$$

In echo-only periods, we want the echo suppression gains to be equal to 0 as it is the case here. Equations 5.92 to 5.93 show that errors in the near-end RTF estimate will not affect the echo suppression performance of the proposed gain rule.

We now consider double-talk periods. In presence of errors in the near-end RTF, the PLDs are written as:

$$\begin{aligned}
\Delta\Phi_{PLD}^{echo}(k,i) =& |\Theta(k,i) + \Delta\Theta(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i) && (5.94)\\
=& |\Theta(k,i) + \Delta\Theta(k,i)|^2 \cdot \left(\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \Phi^{s_1s_1}(k,i)\right)\\
&- \left(|\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + |\Theta(k,i)|^2 \cdot \Phi^{s_1s_1}(k,i)\right)\\
=& \left(|\Theta(k,i) + \Delta\Theta(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\\
&+ \left(|\Theta(k,i) + \Delta\Theta(k,i)|^2 - |\Theta(k,i)|^2\right) \cdot \Phi^{s_1s_1}(k,i) && (5.95)
\end{aligned}$$

$$\begin{aligned}
\Delta\Phi_{PLD}^{near}(k,i) =& \Phi^{y_2y_2}(k,i) - |\Gamma(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) && (5.96)\\
=& \left(|\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + |\Theta(k,i)|^2 \cdot \Phi^{s_1s_1}(k,i)\right)\\
&- |\Gamma(k,i)|^2 \cdot \left(\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \Phi^{s_1s_1}(k,i)\right)\\
=& \left(|\Gamma(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \left(|\hat{\Theta}(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{s_1s_1}(k,i)\\
=& \left(|\hat{\Theta}(k,i)|^2 - |\Gamma(k,i)|^2\right) \cdot \Phi^{s_1s_1}(k,i). && (5.97)
\end{aligned}$$

Rewriting the PLD gain rule with equations 5.95 and 5.97, we obtain:

$$\hat{W}(k,i) = \frac{\Delta\Phi^{near}_{PLD}(k,i)}{\Delta\Phi^{near}_{PLD}(k,i) + \Delta\Phi^{echo}_{PLD}(k,i)} \tag{5.98}$$

$$= \frac{|\Theta(k,i)|^2 - |\Gamma(k,i)|^2}{|\Theta(k,i) + \Delta\Theta(k,i)|^2 - |\Gamma(k,i)|^2} \cdot \frac{\Phi^{s_1 s_1}(k,i)}{\Phi^{s_1 s_1}(k,i) + \Phi^{\tilde{d}_1 \tilde{d}_1}(k,i)}$$

$$= \Omega(k,i) \cdot W(k,i). \tag{5.99}$$

Equation 5.99 shows that errors in the near-end RTF impact on double-talk behavior. The error takes the form of a multiplicative factor Ω. From the expression of Ω, we see that :

- If $\Delta\Theta$ is positive, meaning that $\hat{\Theta}$ is bigger than its real value, Ω will be lower than 1 and the postfilter will apply more echo attenuation than required

$$\Delta\Theta(k,i) > 0 \Longrightarrow \Omega(k,i) < 1 \quad \text{and} \quad \hat{W}(k,i) < W(k,i). \tag{5.100}$$

- In contrasts, if $\Delta\Theta$ is negative, the postfilter might not completely suppress the residual echo

$$\Delta\Theta(k,i) < 0 \Longrightarrow \Omega(k,i) < 1 \quad \text{and} \quad \hat{W}(k,i) > W(k,i). \tag{5.101}$$

It is difficult to state whether the near-end RTF $\hat{\Theta}$ will be over- or under-estimated. We therefore cannot conclude on whether $\hat{W}$ will be overestimated or underestimated.

5.4.3.3 Conclusions on the impact of the RTFs

A new echo suppression gain rule is introduced. In contrast to most existing echo suppression gain rules, the proposed gain rule does not require the estimate of the residual echo PSD. The PSD estimation problem is replaced with the RTFs required for the computation of the new gain rule. Approaches to estimate these RTF are presented in Section 5.4.2. The impact of the estimation errors that might occur in the RTF computation has been analyzed and can be summarized as follows.

1. **During echo-only periods:** The amount of echo suppressed is solely affected by the accuracy of the echo RTF estimate $\hat{\Gamma}$. Errors in the estimation of $\hat{\Gamma}$ lead to a spectral flooring of the echo suppression gain. As a result, the proposed gain rule will achieve less echo suppression and the residual echo might sometimes still be audible after postfiltering. To achieve complete attenuation of the residual echo, we will slightly modify the proposed gain rule:

    - We redefine the proposed gain rule by introducing an overestimation factor. The overestimation factor is a very popular tool used to compensate for PSD estimation error and by so-doing permit to achieve more echo suppression. The gain rule therefore becomes:

    $$\hat{W}(k,i) = \frac{\Phi^{s_1 s_1}(k,i)}{\Phi^{s_1 s_1}(k,i) + \eta \cdot \Phi^{\tilde{d}_1 \tilde{d}_1}(k,i)} \tag{5.102}$$

    $$\hat{W}(k,i) = \frac{\Delta\Phi^{near}_{PLD}(k,i)}{\Delta\Phi^{near}_{PLD}(k,i) + \eta \cdot \Delta\Phi^{echo}_{PLD}(k,i)} \tag{5.103}$$

    where η is the overestimation factor for which we need to find an optimum value.

1. Update of PSDs and cross-PSDs $\Phi^{y_2 y_2}$, $\Phi^{e_1 e_1}$, $\Phi^{x e_1}$ and $\Phi^{x y_2}$
2. Update of RTFs estimate $\hat{\Gamma}$ (Equation 5.65) and $\hat{\Theta}$ (Equations 5.58 or 5.59)
3. Compute the echo and near-end PLDs $\Delta\Phi_{PLD}^{echo}$ and $\Delta\Phi_{PLD}^{near}$ (Equations 5.50 and 5.51)
4. Compute echo suppression gains $\hat{W}$ (Equation 5.103)
5. Gain post-processing based on the DTD $\hat{W}$ (Equation 5.104)

Table 5.1: Summary of the PLD gain update procedure

- In Section 5.3, we introduced a DTD and explained that it could be used to control echo suppression. In our implementation of the proposed gain rule, we use this DTD to achieve more echo suppression during echo-only periods:

  $$\text{if } \Delta\Phi_{PLD}(k,i) < \Phi_{th}(i) \Longrightarrow \hat{W}(k,i) = W_{min} = 0. \tag{5.104}$$

  The minimum gain value W_{min} (i.e. maximum attenuation) is set to 0 in our case because we assume there is no ambient noise. To account for the presence of noise, W_{min} has to be set such as to avoid noise modulation. In the presence of noise, the postfilter should suppress residual echo and output the noise.

2. **During double-talk periods:** The amount of echo suppression is affected by the accuracy of both RTFs. More specifically errors in the echo RTF will result in less echo attenuation (See Equation 5.86). As for the influence of errors in the near-end RTF, we can only state that they lead to errors in the gains and cannot state in advance whether we will over-attenuate or under-attenuate the residual echo.

5.4.4 Summary about the gain rule computation

Table 5.1 summarizes the computation of the proposed PLD gain rule. The first step consists in computing the PSDs and cross-PSDs which we do through auto-regressive smoothing. The near-end RTF $\hat{\Theta}$ can be computed through Equation 5.58 or 5.59. The interest of these two estimates will be assessed in our experiments. The echo RTF is updated during far-end speech activity periods whereas the near-end RTF is updated during near-end only periods. The updated RTFs are used to compute the PLDs $\Delta\Phi_{PLD}^{echo}$ and $\Delta\Phi_{PLD}^{near}$ which are in their turn used to update the echo suppression gains. Prior to the echo suppression itself, the gains are processed according to Equation 5.104 so as to guarantee the maximum attenuation during echo-only periods. The interest of this postprocessing of the gains will be demonstrated in our assessment.

5.5 Performances of the proposed double-talk detector and gain rule

This section reports the performance of the proposed DTD (i.e. Section 5.3) and gain rule (i.e. Section 5.4).

5.5.1 Experimental setup

New DTD and gain rule have been introduced earlier in this chapter: both are based on DM architecture. The performance of the proposed gain rule and DTD are assessed using recorded signals. This assessment also includes comparison of the proposed methods with existing SM echo processing systems.

The echo processing scheme considered is composed of AEC followed by an echo postfilter. The AEC consists of an NLMS algorithm with variable stepsize [Benamar, 1996]. The performance of the three different postfilters considered are assessed in terms of ERLE, SA and of informal listening. The SA as described in Chapter 2 can only be measured if the near-end speech signal $s_1(n)$ is available. With the handset data (i.e. recorded with real devices as presented in Section 5.1.3), $s_1(n)$ is unknown, the microphone signals recorded are already mixed. Nevertheless, the near-end speech signal $s(n)$ at the mouth of the mannequin is known. We assess the DT by computing the SA between the near-end speech signal $s(n)$ from the mouth of mannequin and the output signal of the echo control scheme $\hat{s}_1 n$:

$$SA(k) = 10\log_{10}\frac{\sum_{l=1}^{L} s^2(kL+l)}{\sum_{l=1}^{L} \hat{s}_1^2(kL+l)} \tag{5.105}$$

$$= 10\log_{10}\frac{\sum_{i=1} s^2(k,i)}{\sum_{i=1} \hat{s}_1^2(k,i)} \tag{5.106}$$

$$= 10\log_{10}\frac{\sum_{i=1} s^2(k,i)}{\sum_{i=1} (G_1(k,i)\cdot W(k,i))^2 s^2(k,i)} \tag{5.107}$$

The modified SA as defined in Equation 5.107 does not only account for the SA due to postfiltering but will also account for the level difference due to the acoustic path between the mouth and the primary microphone $g_1(n)$. In this particular case, the absolute SA values will not really have a meaning: the acoustic path $g_1(n)$ can cause the SA to be artificially very high or very low. The modified SA used for the handset data nevertheless gives an information about the relative SA between the different schemes considered and tells us which system achieves the most or least SA.

The assessment of the PLD based methods is performed in three main steps:

- **Assessment of the PLD based DTD:** As explained in Section 5.3.2, the PLD based DTD is compatible with any single microphone echo processing scheme. The proposed DTD is implemented and assessed within the echo postfilter used in Chapter 3; echo postfilter which is applied to primary microphone only. Since, we are focus on the DTD itself there is no need to use it with the secondary microphone: the conclusions on the DTD behavior will be independent of the microphone and will be the same no matter the microphone signal used.We use the DTD to postprocess the estimate of the residual echo PSD according to Equations 5.42 and 5.43 as follows:

$$\begin{aligned} \text{if } \Delta\Phi_{PLD}(k,i) < \Phi_{th} \Longrightarrow \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) &= \eta_{echo}\cdot\Phi^{e_1e_1}(k,i) \\ \text{otherwise } \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) &= \eta_{dt}\cdot\Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \end{aligned} \tag{5.108}$$

  In our experiments, we analyze the impact of the value of the threshold Φ_{th} which we consider to be the same for all frequency bins i. For a given frequency bin i and frame k, the overestimation factor is set to $\eta_{echo} = 8$ if echo-only is detected or to $\eta_{dt} = 4$ if DT is detected. These values were proven to achieve good echo suppression in echo-only periods and an acceptable level of distortion during DT periods. In the assessement, the baseline echo postfilter is denoted Baseline SM and its improved version is denoted SM - DTD. This set of experiments is reported in Section 5.5.2.

- **Assessment of the PLD based gain rule:** The performance of the new gain rule is also assessed and compared to that of the baseline SM postfilter. We show in Section 5.4.3 that errors in the RTF might result in low echo suppression during periods of echo-only. We explain in Section 5.4.3 that an overestimation factor could be used to increase the amount of echo suppression and redefine the proposed gain as follows:

$$\hat{W}(k,i) = \frac{\Delta\Phi_{PLD}^{near}(k,i)}{\Delta\Phi_{PLD}^{near}(k,i) + \eta \cdot \Delta\Phi_{PLD}^{echo}(k,i)}, \tag{5.109}$$

  where Φ_{PLD}^{near} and Φ_{PLD}^{echo} are computed according to Equations 5.50 and 5.51. Part of our experiments aims to determine an optimum value for this overestimation factor η. The values of η used here are different from those of η_{echo} and η_{dt} used for the SM postfilter. In addition, in our implementation of the DM postfilter η has the same value during echo-only and double-talk periods. Later on in the experiments the DTD is used to improve even further the echo suppression by postprocessing the PLD gain as follows:

$$\text{if } \Delta\Phi_{PLD}(k,i) < \Phi_{th} \Longrightarrow W(k,i) = W_{min}. \tag{5.110}$$

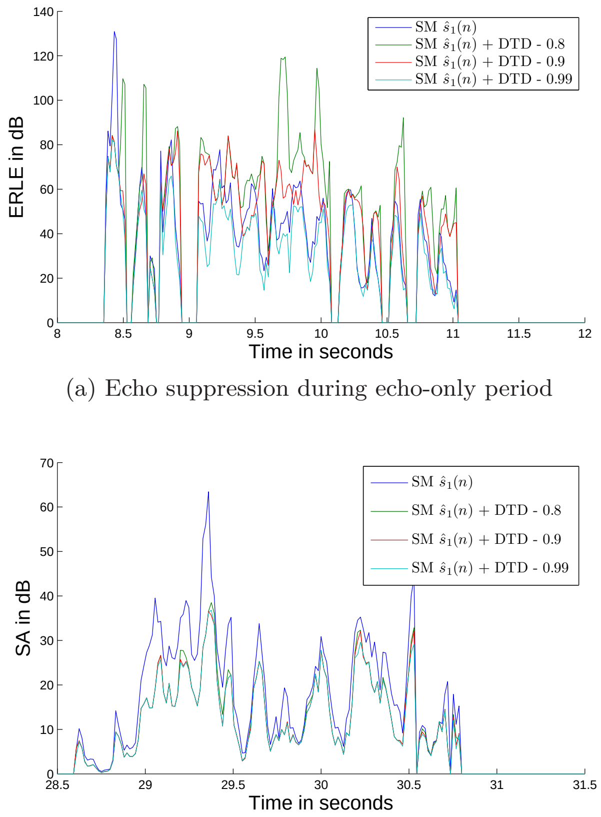
  The proposed gain rule requires the knowledge of the near-end RTF for which we propose two different estimates. The impact of these two estimates is also part of our assessment. In our experiments, the PLD gain is denoted DM. This set of experiments allows us to fully assess the new gain rule. Our observations are reported in Section 5.5.4.

- **Overall assessment of the PLD based DTD and gain rule:** The impulse responses measured with the mock-up phone are used to generate a test dataset. This dataset is used to assess our proposed DM methods (DTD and gain rule) to a state-of-the-art DM echo processing system. This experiments are discussed further in Section 5.5.4. Section 5.5.5 reports experiments performed with data from real device. Finally, a synthesis of our experiments through subjective quality assessment is presented in Section 5.5.6.

5.5.2 Influence of the threshold value on the DTD

Figure 5.9 shows the ERLE and SA curves for the SM postfilter used with the PLD. More specifically different values of thresholds are applied to the PLD values as specified in Equation 5.108. Three different threshold values of Φ_{th} are considered: -0.8, -0.9 and -0.99.

At the beginning of the echo-only period, the use of the DTD does not lead to greater ERLE than the baseline SM $\hat{s}_1(n)$ case. This is due to the transient period of the PSDs involved in the computation of the PLD. The transient time refers to the time an estimate to reach its optimum. These PSDs are computed through autoregressive smoothing and need some time to rise after an increase in the level of the error. During this transient time, the residual echo PSD will be over- or under-estimated depending on whether the level of the error signal is rising or falling. Once the error PSD reaches its steady state, the



(a) Echo suppression during echo-only period

(b) Speech attenuation during double-talk period

Figure 5.9: ERLE and SA for different values of threshold on the PLD values. Impulse responses used are measured in office environment

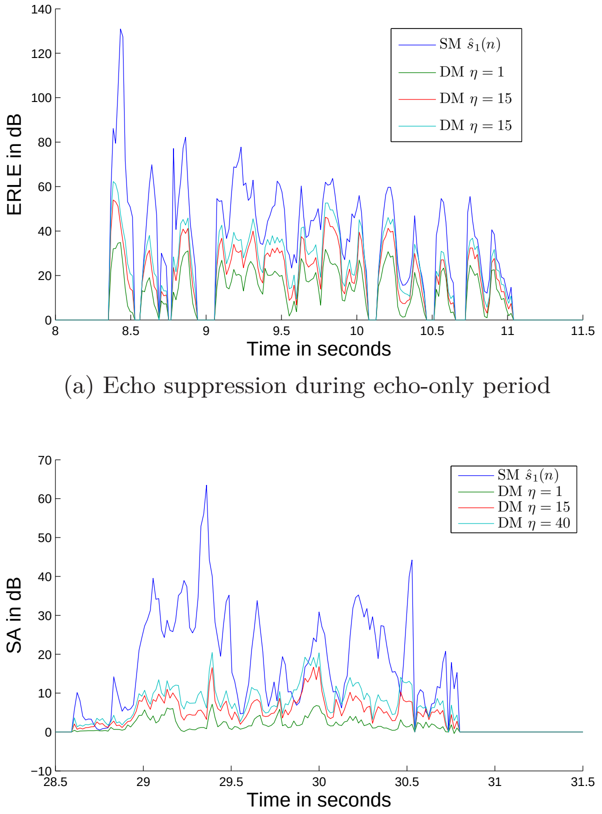
use of DTD with appropriate values of Φ_{th} permits to improve the ERLE performance. When the threshold Φ_{th} is set to -0.99, the proposed system achieves less echo suppression than the baseline system whereas for the other values of Φ_{th} (-0.8 and -0.9), significant improvement can be observed. In general, Figure 5.9 (a) shows that the higher Φ_{th}, the more echo suppression we achieve.

The SA curve in Figure 5.9 (b) shows that the use of the DTD permits to slightly reduce attenuation of the near-end speech signal during double-talk periods. This SA improvement is due to the use of the two-step overestimation controlled with the DTD which results in lower overestimation factor during DT. Figure 5.9 (b) shows that the value of the threshold does not significantly impacts the SA.

In the reminder of our experiments, when using the DTD, Φ_{th} will be set to -0.9 which was found to give the best performance for our system.

5.5.3 Assessment of the PLD based gain rule

Figure 5.10 (a) shows the amount of echo suppression achieved by the proposed PLD gain rule for different values of the overestimation factor η (Equation 5.109). We observe that for $\eta = 1$ the proposed system does not achieve as much echo suppression as the baseline SM $\hat{s}_1(n)$ system: this poor behavior is the direct consequence of errors in the echo RTF. As expected the amount of echo suppression increases as the overestimation factor increases.



(a) Echo suppression during echo-only period

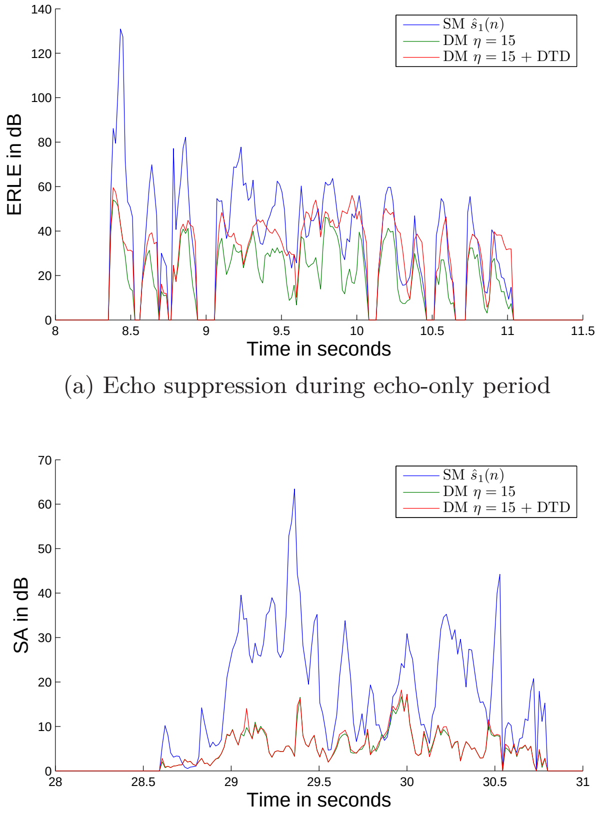
(b) Speech attenuation during double-talk period

Figure 5.10: Influence of the overestimation factor within the PLD gain rule. Impulse responses used are measured in office environment

The SA curve in Figure 5.10 (b) shows that the increase of the overestimation factor η results in a SA increase. In Section 5.4.3, we show that errors in the echo and near-end RTF estimates are responsible for the SA during DT periods. Figure 5.10 (b) shows that, regardless of these errors, the PLD gain rule still leads to very low SA values compared to the baseline SM $\hat{s}_1(n)$ system. Although it achieves less echo suppression, the PLD gain rule remains an appealing alternative gain rule for its low SA.

The PLD based DTD can be used to achieve further echo suppression. Figure 5.11 shows the performance of the PLD gain when used with the DTD according to Equation 5.110. We observe from Figure 5.11 (a) that the use of the DTD permits to increase the amount of echo suppression: an improvement of up to 10dB can be observed in some regions. The poor performance observed between between time $t = 8.4$s and 9s is mainly due to the transient time of PSDs and RTF. The computation of the PLD gain rule requires the estimation of the RTFs which are computed with smoothed PSDs. In addition, in this same region, the PSDs involved in the computation of the normalized PLD have not yet reached their optimum and therefore influences the DTD. The SA curve in Figure 5.11 (b) shows that the use of the DTD marginally affects the performance of the PLD gain rule during DT periods. In the reminder of our experiments, the PLD gain rule is implemented with the DTD.

Until now, the PLD gain rule was implemented using the MMSE estimate of the



(a) Echo suppression during echo-only period

(b) Speech attenuation during double-talk period

Figure 5.11: Usage of the DTD with the PLD gain rule. Impulse responses used are measured in office environment

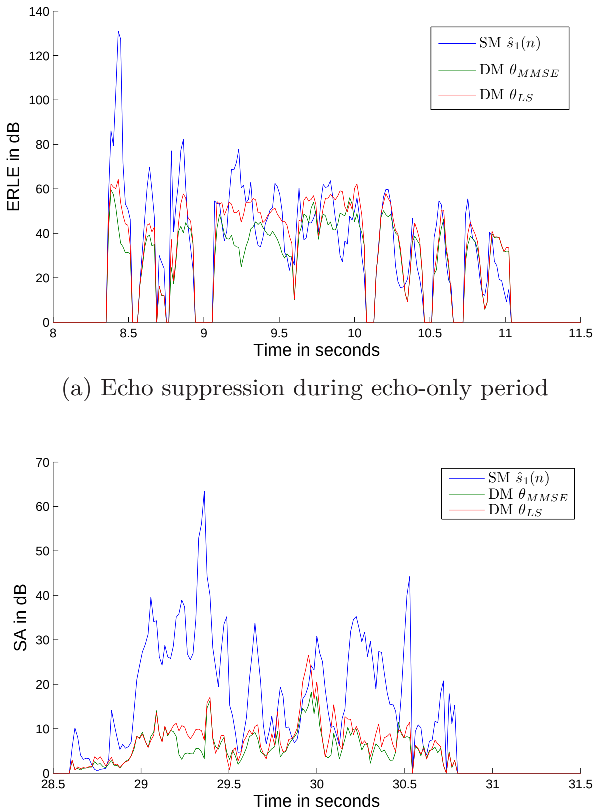
near-end RTF. In Section 5.4.2.1, we introduced two estimates of the near-end RTF. Figure 5.12 (a) and 5.12 (b) illustrate the impact of these estimates on the performance of the postfilter. As illustrated in Figure 5.12 (a), the use of the unbiased estimate $\hat{\Theta}_{LS}$ sometimes permits to achieve more echo suppression. Unfortunately, as we can see in Figure 5.12 (b) it also results in an increased SA compared to that of the MMSE estimate $\hat{\Theta}_{MMSE}$. However the SA introduced with the unbiased estimate $\hat{\Theta}_{LS}$ remains largely lower than that introduce by the Baseline SM $SM\hat{s}_1(n)$ system.

The LS estimate leads to better echo suppression whereas the MMSE estimate introduces less distortion during double-talk. In the remaining of our experiments, we choose to use the MMSE estimate for its computational simplicity.

5.5.4 Performance with data from mock-up phone

In this section, we present a more global analysis of our systems using a large database of speech signals. Impulse responses measured with the handsfree mock-up phone are used to generate a database of test signals with SER ranging from -5dB to 10dB at the primary microphone. Our DM approaches are assessed in this comparison which includes 3 differents systems:

- Since, it is difficult to find a consistent DM echo processing in the litterature for comparison with our system, we design a baseline DM approach based on the com-



(a) Echo suppression during echo-only period

(b) Speech attenuation during double-talk period

Figure 5.12: Influence of near-end RTF estimate. Impulse responses used are measured in office environment

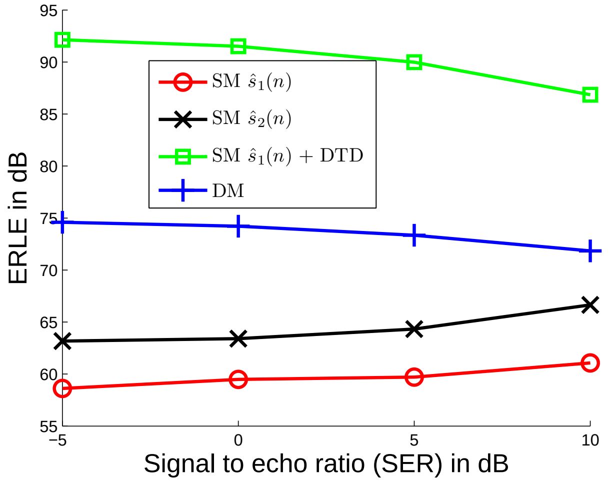
parison of the output signal of SM echo processing scheme to each microphone path. Therefore, this baseline approach outputs two signals $\hat{s}_1(n)$ and $\hat{s}_2(n)$ which we denote $SM\hat{s}_1(n)$ and $SM\hat{s}_2(n)$ respectively. In a real-time implementation, one would need to choose between both outputs: such a choice can be done in real-time using an additional decision block or offline (i.e. when designing our echo processing system). In our case, we use both outputs as reference system for comparison with our system. Advantages and inconveniences of each output will be discussed.

- We use our DTD as in Section 5.5.2 to control a SM echo postfilter. In this case, we apply the echo processing on the primary microphone and denote this output $SM\hat{s}_1(n) - DTD$: this output can be seen as an improved version of the baseline SM $s_1(m)$ system.

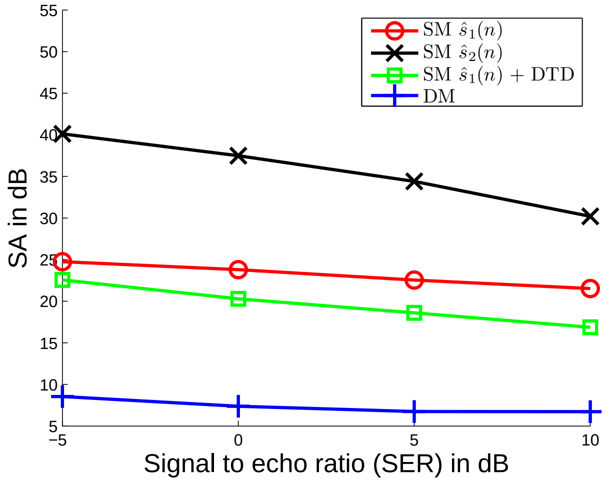
- Lastly, our new gain rule as tuned in Section 5.5.3 is considered in our experiments. We denote this method DM.

5.5.4.1 Performance of proposed DTD and gain rule

Figures 5.13 (a) and 5.13 (b) show the ERLE and SA as functions of the SER. The baseline SM $s_2(n)$ achieves less echo suppression than the baseline SM $s_2(n)$ but achieves less SA during DT periods. The difference in performance between both systems is due



(a) Echo suppression



(b) Speech attenuation

Figure 5.13: Objective performance against SER

to transducers arrangement on the mock-up phone. The secondary microphone is closer to the loudspeaker than the primary. As a result, the secondary microphone will capture more echo than the primary microphone. This will lead to high echo suppression but also strong distortions of the near-end speech signal for the secondary microphone. At some point, the baseline SM $s_1(m)$ offers a better trade-off between the amount of echo suppression and distortions during DT periods than baseline SM $s_2(m)$

In terms of ERLE, we observe that the use of the DTD permits to significantly improve the performance of the baseline SM $s_1(n)$ solution: an improvement of more than 15dB is observed. The ERLE curves in Figure 5.13 (a) also shows that the DM achieve slightly less echo suppression than the baseline SM $s_1(n)$ solution: there is a difference of about 4dB. The SA curves in Figure 5.13 (b) shows that the PLD gain rule achieves the best performance. Although the DM does not achieve as much echo suppression that the other methods, it remains very interesting for its double-talk behavior. The use of the

DTD in the baseline SM $s_1(m)$ system permits to reduce its SA of about 1-3dB. This SA improvement can be explained by the two-step overestimation factor which is used and controlled by the DTD.

5.5.4.2 Impact of the position of the secondary microphone

We mentioned in Section 5.3 that placing the secondary microphone closer to the loudspeaker could improve the performance of our DTD. The recorded impulse response are used to generate a dataset of signals accounting for different distances between the loudspeaker and the secondary microphone. It is generally admitted that in non-reverberant environment as it is case in our recording, the amplitude of the min delay of an acoustic path is inversely proportional to the distance between the source and the microphone [Habets, 2007]. The distance $d^0_{xy_2}$ between the loudspeaker and the secondary microphone of our mock-up is about 5cm. We simulate two other positions of the secondary microphone d_{xy_2} (2cm and 1cm) by simply amplifying the impulse response $h_2^{mod}(n)$ proportionally to the inverse of the target distance d_{xy_2}:

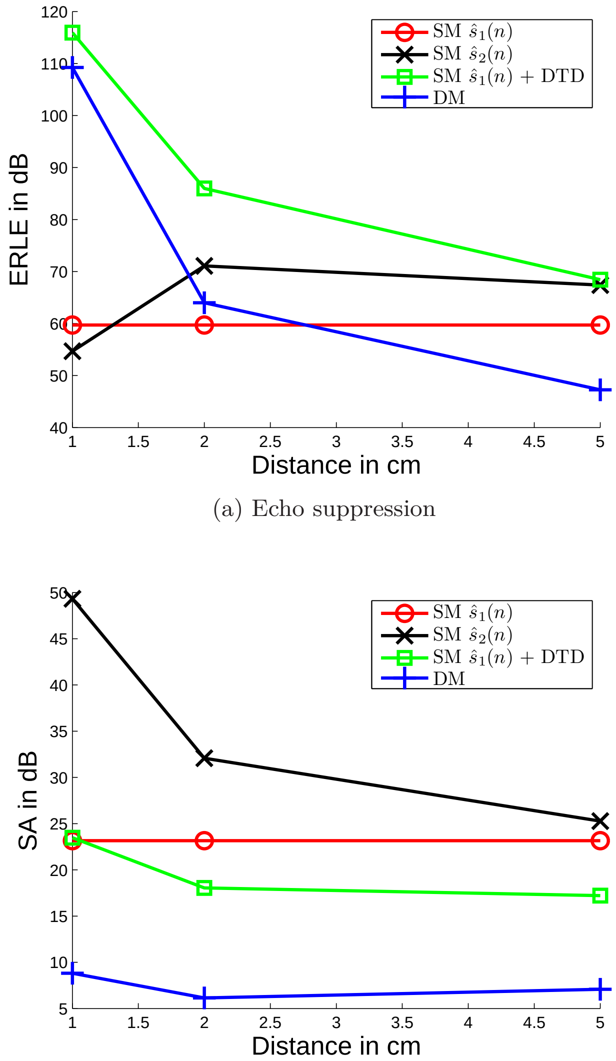
$$h_2^{mod}(n) = \frac{d^0_{xy_2}}{d_{xy_2}} \cdot h_2(n). \tag{5.111}$$

The distance d_{sy_2} between the mouth of the mannequin and secondary microphone is equal to 30 or 50cm. Moving the secondary microphone for 5cm as it is the case here will have a significant impact on $g_2(n)$. For this reason, we assume that moving the secondary microphone towards the loudspeaker will only modify the echo path: $h_2(n)$ then becomes $h_2^{mod}(n)$. The amplified impulse response $h_2^{mod}(n)$ is used in conjunction with the measured impulse responses $h_1(n)$, $g_1(n)$ and $g_2(n)$ to generate a dataset of test signals. The resulting dataset contains microphone signals with SER which we set ranging from -5dB to 10dB for the primary microphone.

Figures 5.14 (a) and 5.14 (b) show the ERLE and SA performance as functions of the distance between the loudspeaker and the secondary microphone. Once more the baseline SM $s_1(n)$ system gives a better tradeoff between echo suppression and distortion during DT periods than the baseline SM $s_2(n)$. Performance of the baseline SM $s_2(n)$ decreases (reduction of the amount of echo suppression and increase of amount of distortion in DT) as d_{xy_2} decreases. The poor echo attenuation for small values of d_{xy_2} is due to that the more the secondary microphone gets closer to the loudspeaker, the more clipping might occur on this microphone. As a consequence, the baseline SM $s_2(n)$ lead to poor echo suppression due to the presence of clipping (which is seem as non-linearity in our echo processing scheme).

As the secondary microphone is moved closer to the loudspeaker the amount of echo suppression achieved by the DM solutions (SM combined with DTD and PLD gain rule) increases. The gap between the DM and the SM $s_1(n)$ + DTD systems reduces as the secondary microphone gets closer to the loudspeaker. For all DM solutions considered, the SA increases as the distance d_{xy_2} decreases. As we can see in Figure 5.14 (b), the DM system clearly outperforms the other methods in terms of SA: such differences will clearly be noticeable during a conversation. The PLD gain rule will potentially result in better speech quality, better intelligibility.

As the secondary microphone gets closer to the loudspeaker, the more saturation might occur on the secondary microphone. Experiments show that saturations (i.e. clipping) of the secondary microphone do not altered the performance of the DM solutions (SM



(a) Echo suppression

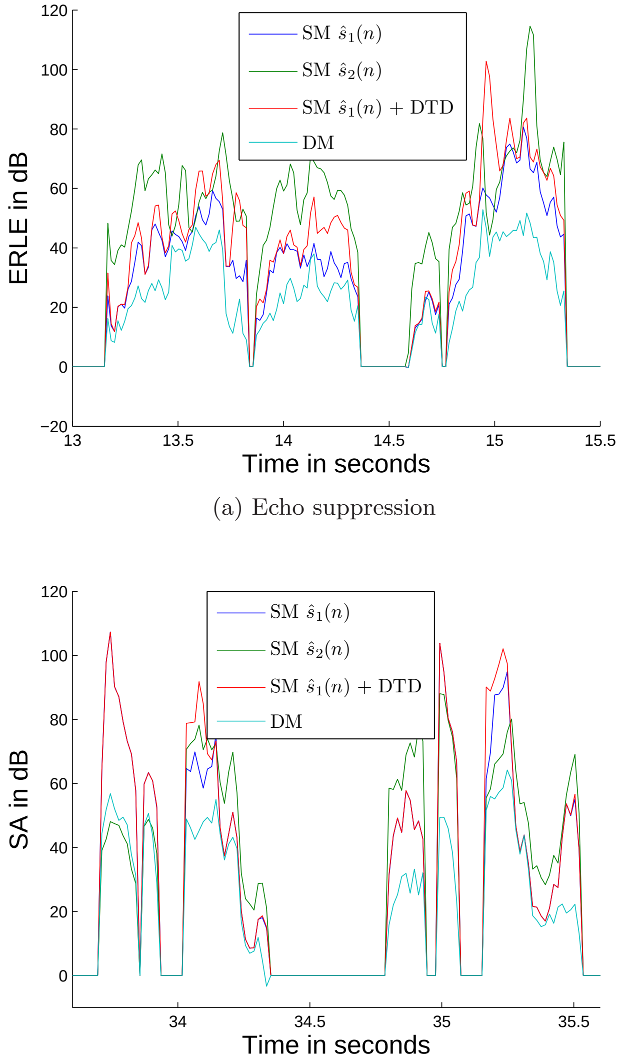
(b) Speech attenuation

Figure 5.14: Objective performance against distance between loudspeaker and secondary microphone

combined with DTD and PLD gain rule). The experimentation about the position of the secondary microphone can be interpreted as a recommendation about the positioning of the transducers on the mobile devices. Optimum behavior of the DM solutions are obtained by considering some hardware constraints:

- placing the microphones in bottom-top configuration
- Placing the secondary microphone as closed as possible of the loudspeaker.

Referring the ERLE curve in Figure 5.14 (a), we see that if d_{xy_2} is lower than 2.75cm, the PLD gain outperforms the Baseline SM solution. This value of d_{xy_2} might be interpreted as the maximal distance between to loudspeaker and the microphone which guarantees that the PLD gain rule achieves more echo suppression than the Baseline solution.



(a) Echo suppression

(b) Modified SA

Figure 5.15: Echo suppression performance with real data

In real time systems, the echo postfilter uses a noise floor so as to avoid excessive attenuation during echo-only periods and noise modulation. The noise floor is used to force the output of the postfilter to be higher or equal to the noise level. The use of the noise floor improves the overall perceived quality of processed signals but results in lower echo suppression. Therefore, although the PLD gain rule does achieve the most ERLE, it still be considered an alternative echo postfiltering approach. In addition it presents the advantage of reduced level of near-end speech attenuation during DT periods especially when the secondary microphone is placed close to the loudspeaker.

5.5.5 Assessment with handset recording

Figure 5.15 (a) shows the ERLE curves measured with recording from real mobile device (see Section 5.1.3). We observe that the baseline SM and the improved SM systems have similar performance. The PLD gain rule achieves slightly less echo reduction compared to the baseline SM $\hat{s}_1(n)$ and improved SM system. We nevertheless see that the PLD gain rule can achieve significant echo attenuation. It is interesting to note that echo suppression with real recordings are consistent with the observation made for handsfree.

We explained in Section 5.5.1 that for the data recorded with the real devices, the reference near-end signal $s_1(n)$ required for the computation of the SA is unavailable. The SA curve reported here is computed according to Equation 5.107. Figure 5.15 (b) shows an example of modified SA curves. This time the baseline SM $\hat{s}_1(n)$ seems to achieve more SA than the baseline SM $\hat{s}_2(n)$. Since the SA reported in this curve also account for the acoustic between the mouth and each of the microphone (See Equation 5.107), it might be difficult to compare these two SA. Moreover, in Section 5.1.3 we explain that the $G_2(k,i) << G_1(k,i)$. Therefore part of the SA difference between the baseline SM $\hat{s}_1(n)$ and baseline SM $\hat{s}_2(n)$ is due to the difference between $G_2(k,i)$ and $G_1(k,i)$. The modified SA of the baseline SM $\hat{s}_2(n)$ system cannot be reliably be compare to the other systems.

Nevertheless, we see that the baseline SM $\hat{s}_1(n)$ system achieves the most SA whereas the PLD gain rule achieves the least SA. We also see that the improved SM system outperforms the baseline SM $\hat{s}_1(n)$ system: the SA improvement sometimes reaches 10dB. Once more, we note the consistency between our observations with realistic data and our observations using the data measured with the mock-up phone.

5.5.6 Informal listening tests

Informal listening show that signals processed by the PLD gain rule sound slightly reverberant-like. We note the presence of musical noise signals processed through the baseline SM $\hat{s}_1(n)$ and the baseline SM $\hat{s}_1(n)$ + DTD systems. The use of the PLD methods result in better intelligibility of the near-end speech. This intelligibility improvement is clearly audible for the PLD gain rule: low level signals and transitions between syllables are better conserved. Speech quality and artifacts perceived are the same for the handset and handsfree signal.

The use of the PLD method (DTD and/or gain rule) is an appealing alternative to improve echo suppression for handsfree scenarios. The DTD is a simple add-on module which permits to significantly improve the echo suppression of the baseline echo processing scheme. From an industrial point of view, our results show that the DM hardware constraint can be translated into significant improved echo suppression at the cost of a very low computational complexity increase. The PLD gain rule is a more radical solution to echo suppression as it requires a completely new implementation. However, its DT performance makes it a good target for long-term development.

5.6 Conclusions

In this chapter, we introduced the problem of acoustic echo in DM terminals. We analyzed the echo problem based on recordings with a mock-up phone and a real mobile device. In both cases, the transducers are placed in bottom-top configuration. This configuration of the transducers leads to an important level difference between the microphone signals in echo-only periods. A DTD that exploits this level difference is introduced.

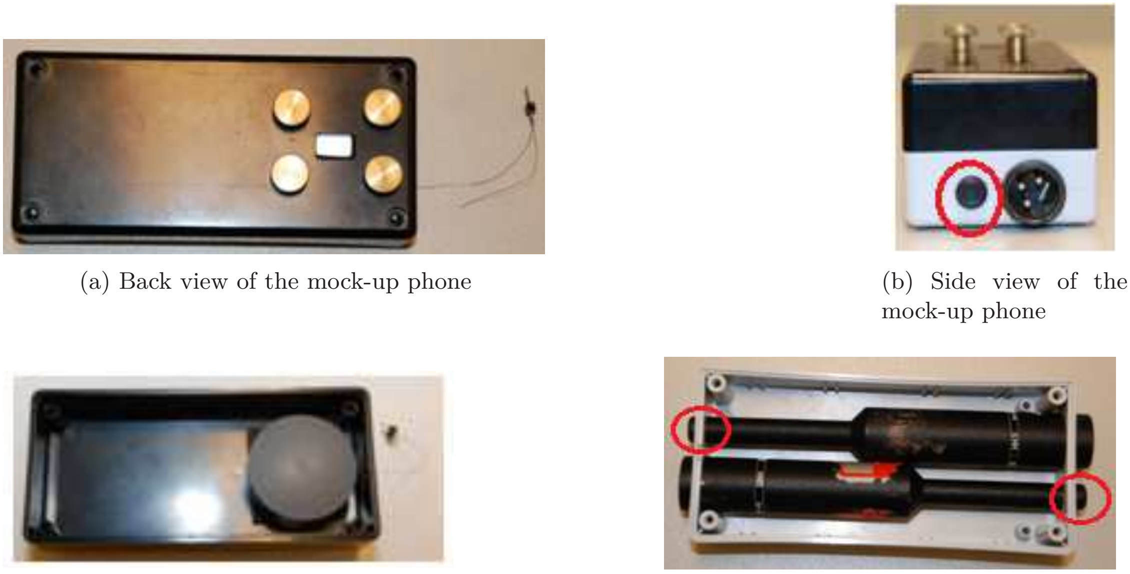
We propose to tackle the echo problem using adaptive echo cancellation followed by postfiltering. We explain why state-of-the-art adaptive filters can still be used. The postfilter on its side uses the DM signals to output an estimate of the near-end speech signal. We proposed a novel DM gain rule which uses two RTFs. Mathematical analysis shows that errors in the echo RTF estimation lead to insufficient echo suppression in echo-only periods and DT periods whereas errors in the near-end RTF only impact on DT performances. To limit the impact of RTF estimation errors, the proposed gain rule is implemented in combination with our DTD.

For assessment, the proposed DTD is implemented within the echo postfilter used in Chapter 3. This experiment shows that its use permits to increase the amount of echo suppressed in echo-only periods while reducing the amount of distortion during DT periods. Assessment of the proposed DM gain rule showed that introduces even less distortion than the Baseline SM echo postfilter of Chapter 3 and its improved version.

The DTD and gain rule proposed in this chapter are based on the level difference between the microphone signals. Our simulations show that the performance of the proposed DTD and gain rule increases as this level difference increases. From these simulations, we show that there is a potential for hardware recommendations about the configuration of the transducers on the devices:

- the primary microphone should be placed at the bottom of the device such that in handset, it maximally picks the near-end speech.

- the secondary microphone should be placed at the top of devices such as to be the further away from the primary microphone. With such arrangement, we guarantee good performance of the proposed methods: clear distinction between echo-only and double talk periods using the normalized PLD and efficient echo suppression using the PLD gain rule.



(a) Back view of the mock-up phone

(b) Side view of the mock-up phone

(c) Loudspeaker: inside installation

(d) Loudspeaker: inside installation

Figure 5.16: Mock-up phone: loudspeaker installation

Length	12.8
Width	5.8
Height	4.6
Distance between microphone	13

Table 5.2: Dimensions of mock-up phone in cm

5.A Recording setup

5.A.1 Description of mock-up phone

The mock-up phone used in our experiments is illustrated in Figure 5.16. It basically consists of a plastic box in which a loudspeaker (15x11x3.5 MM RA SPEAKER) and two MM1 microphones are installed. The dimension of the mock-up phone are specified in Table 5.2. Except for its height, the dimensions of our mock-up are similar to that of some devices that are on the market today (Samsung Galaxy S4). In order to be able to place the microphone inside the box, the height of the box could not be reduced further.

The loudspeaker used permits to simulate handsfree use-case. The model chosen (15x11x3.5 MM RA SPEAKER) is one which already used for some mobile devices.

5.A.2 Signal recording setup

As mentioned in Section 5.1.2, our mock-up is used to record signals in three different environments: cabin, office and meeting room. In the cabin environment, the phone is placed at 30 cm straight in front of the artificial head's mouth: the phone and the mouth are approximately in the same plane. This setup is illustrated in Figure 5.17. In the office and meeting room, the phone is placed on a table 50cm away from the mouth of the mannequin according to the ITU-T P340 recommendation [ITU-T, 1996a]. For all setups mentioned here, impulse responses measured have a length of 100ms which is long enough sinc we are not in reverberant acoustic environments.

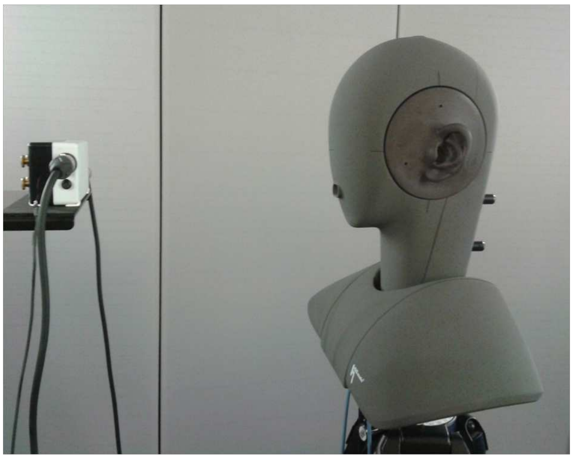


Figure 5.17: Illustration of our recording setup in the cabin environment

Chapter 6

Dual microphone based echo postfilter

Contents

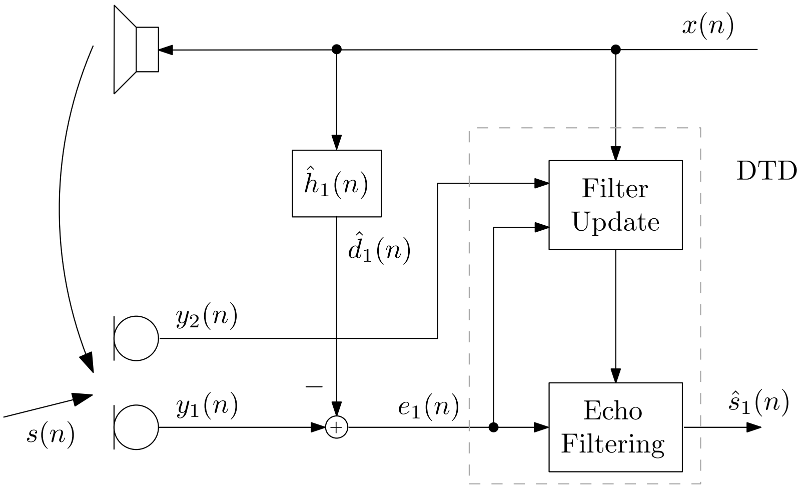


Figure 6.1: Echo processing scheme for DM mobile terminals

In this chapter, we continue our focus on the problem of echo suppression for DM terminals. The echo processing scheme considered is still that described in Chapter 5 i.e. composed of adaptive filtering followed by DM postfiltering. In this Chapter, we focus more specifically on the echo postfilter for which a new estimate of the residual echo PSD is proposed. The echo PSD estimator is based on the correlation between microphone signals. This correlation based logic is extended so as to address the problem of loudspeaker non-linearity. The proposed PSD estimates are assessed using the same dataset used in Chapter 5.

The methods presented in this Chapter have been patented [**Yemdji** et al., 2012a] and published in [**Yemdji** et al., 2012c].

6.1 Residual echo power spectrum estimate for dual microphone devices

6.1.1 Echo power spectrum estimate

The echo processing scheme of interest is that of Chapter 5 which is shown in Figure 6.1 for reminder. As explained previously, even though we consider DM devices, echo suppression gain can still be computed with existing SM echo suppression gain rules. Let us consider the Wiener gain as computed in Chapter 3, i.e. Wiener gain rule with SER computed through decision directed approach (DDA):

$$W(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)} \quad \text{with} \quad \xi(k,i) = \frac{\Phi^{s_1 s_1}(k,i)}{\Phi^{\tilde{d}_1\tilde{d}_1}(k,i)} \tag{6.1}$$

where $\xi(k,i)$ denotes the SER at the primary microphone after AEC. As explained in Chapter 3, the computation of the SER through the DDA only requires the knowledge of the residual echo PSD for which we propose a new estimate.

The residual echo PSD is defined as in Chapter 5:

$$\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) = |\tilde{H}_1(k,i)|^2 \cdot \Phi^{xx}(k,i). \tag{6.2}$$

We saw in Chapter 5 that using the echo and near-end RTFs,

$$\Gamma(k,i) = \frac{H_2(k,i)}{\tilde{H}_1(k,i)} \quad \text{and} \quad \Theta(k,i) = \frac{G_2(k,i)}{G_1(k,i)}, \tag{6.3}$$

the microphone signal PSDs can be written as:

$$\Phi^{e_1e_1}(k,i) = \Phi^{s_1s_1}(k,i) + \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) \tag{6.4}$$

$$\Phi^{y_2y_2}(k,i) = \Phi^{s_2s_2}(k,i) + \Phi^{d_2d_2}(k,i) \tag{6.5}$$

$$= |\Theta(k,i)|^2 \cdot \Phi^{s_1s_1}(k,i) + |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \tag{6.6}$$

An estimate of the residual echo PSD can be derived from the expressions of the microphone PSDs in Equations 6.4 and 6.6:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\Theta(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i)}{|\Theta(k,i)|^2 - |\Gamma(k,i)|^2}. \tag{6.7}$$

The required RTFs in Equation 6.7 are the same as those used in Chapter 5. They can be computed as described in Section 5.4.2. The residual echo PSD estimate is therefore written as:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\hat{\Theta}(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i)}{|\hat{\Theta}(k,i)|^2 - |\hat{\Gamma}(k,i)|^2}. \tag{6.8}$$

The residual echo PSD is only computed during far-end speech activity periods, meaning during both echo-only and DT periods. During other periods, the residual echo PSD is set to zero. This permits to avoid PSD mis-estimations that might occur during near-end or noise-only periods.

6.1.2 Analysis of the proposed DM PSD estimate behavior

The proposed PSD estimate is directly dependent on the RTFs estimate. It is of interest to understand how these RTF estimates might affect the PSD estimate behavior. The residual echo PSD is estimated during periods of far-end speech activity. Echo-only and DT periods are considered for this analysis. Similarly as in Chapter 5, we model the echo and near-end RTF estimation errors as an additive noise and analyze their impact separately.

6.1.2.1 Impact of the near-end RTF

Let us express the near-end RTF estimate as follows:

$$\hat{\Theta}(k,i) = \Theta(k,i) + \Delta\Theta(k,i) \tag{6.9}$$

where $\Delta\Theta(k,i)$ represents the error in the near-end RTF. By introducing Equation 6.9 into the expression of the residual echo PSD estimate of Equation 6.8, we obtain the following:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\Theta(k,i) + \Delta\Theta(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i)}{|\Theta(k,i) + \Delta\Theta(k,i)|^2 - |\hat{\Gamma}(k,i)|^2}. \tag{6.10}$$

If we consider echo-only periods, the microphone signal PSDs can be written as:

$$\Phi^{e_1e_1}(k,i) = \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) \tag{6.11}$$

$$\Phi^{y_2y_2}(k,i) = \Phi^{d_2d_2}(k,i) = |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \tag{6.12}$$

With the use of Equations 6.11 and 6.12, the residual echo PSD estimate now becomes:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\Theta(k,i)+\Delta\Theta(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) - |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)}{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\hat{\Gamma}(k,i)|^2} \tag{6.13}$$

$$= \frac{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\Gamma(k,i)|^2}{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\hat{\Gamma}(k,i)|^2} \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) = \Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \tag{6.14}$$

Equation 6.14 shows that estimation errors that occur in the near-end RTF will not really impact on the residual echo PSD estimate. In a real-time implementation, the near-end RTF vector will be initialized to a certain value. Without any a-priori information about the acoustic environment, the initial value of $\hat{\Theta}$ will be completely different from the real near-end RTF $\hat{\Theta}$, meaning that the RTF estimation error $\Delta\Theta$ will be high. Equation 6.14 tells us that, even if the near-end RTF is poorly initialized, so long as the initial value is not null, the residual echo PSD estimate during echo-only will still be correct.

If we now consider double-talk periods and rewrite the residual echo PSD estimate using Equations 6.4 and 6.6, we obtain the following:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\Theta(k,i)+\Delta\Theta(k,i)|^2 \cdot (\Phi^{s_1s_1}(k,i) + \Phi^{\tilde{d}_1\tilde{d}_1}(k,i))}{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\hat{\Gamma}(k,i)|^2} - \frac{(|\Theta(k,i)|^2 \cdot \Phi^{s_1s_1}(k,i) + |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i))}{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\hat{\Gamma}(k,i)|^2} \tag{6.15}$$

$$= \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \frac{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\Theta(k,i)|^2}{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\hat{\Gamma}(k,i)|^2} \cdot \Phi^{s_1s_1}(k,i) \tag{6.16}$$

$$= \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \Delta\Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \tag{6.17}$$

The term $\Delta\Phi^{\tilde{d}_1\tilde{d}_1}$ denotes the residual echo PSD estimation error. Equations 6.15 to 6.17 show that errors in the near-end RTF estimate directly impact on the DT behavior of the proposed residual echo estimate. From Equation 6.16, we see that the error term $\Delta\Phi^{\tilde{d}_1\tilde{d}_1}$ depends of the near-end speech component. This error term $\Delta\Phi^{\tilde{d}_1\tilde{d}_1}$ will result in both over- and under-estimation of the residual echo PSD. Moreover, if we consider that the near-end and far-end speech signals do not have the same pitch frequency (i.e. echo and near-end speech signals have different spectral components), this dependence on the near-end speech component means that some frequency components of the near-end speech signal will be introduced into the residual echo PSD estimate. We can state that the near-end RTF estimation error will be responsible for part of the near-end speech attenuation during DT periods. If we consider a phone conversation which starts with a DT period, meaning that the near-end RTF has not yet been updated, then DT behavior will be significantly affected.

6.1.2.2 Impact of the echo RTF

Similarly as for the near-end RTF, we analyze the impact of the echo RTF on the residual echo PSD accuracy. Let us denote the echo RTF estimation error as $\Delta\Gamma(k,i)$ such that:

$$\hat{\Gamma}(k,i) = \Gamma(k,i) + \Delta\Gamma(k,i). \tag{6.18}$$

With Equation 6.18, the residual echo PSD estimate becomes:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\Theta(k,i)|^2 \cdot \Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i)}{|\Theta(k,i)|^2 - |\Gamma(k,i)+\Delta\Gamma(k,i)|^2}. \tag{6.19}$$

Considering echo-only periods, the residual echo estimate PSD is then given by:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\Theta(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) - |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)}{|\Theta(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2} \tag{6.20}$$

$$= \frac{|\Theta(k,i)|^2 - |\Gamma(k,i)|^2}{|\Theta(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2} \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) \tag{6.21}$$

$$= P(k,i) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \tag{6.22}$$

The lower the RTF estimation error, the closer the multiplicative term in Equation 6.22 $P(k,i)$ is to 1:

$$|\Delta\Gamma(k,i)| \to 0 \Rightarrow |P(k,i)| \to 1. \tag{6.23}$$

Equations 6.20 to 6.22 tell us that errors which occur in the computation of the echo RTF create some over- or underestimation of the residual echo.

Now considering DT periods, the residual echo PSD estimate becomes:

$$\begin{aligned}\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) &= \frac{|\Theta(k,i)|^2 \cdot \left(\Phi^{s_1s_1}(k,i) + \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\right)}{|\Theta(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2} \\ &- \frac{\left(|\Theta(k,i)|^2 \cdot \Phi^{s_1s_1}(k,i) + |\Gamma(k,i)|^2 \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i)\right)}{|\Theta(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2}\end{aligned} \tag{6.24}$$

$$= \frac{|\Theta(k,i)|^2 - |\Gamma(k,i)|^2}{|\Theta(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2} \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) \tag{6.25}$$

$$= P(k,i) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i). \tag{6.26}$$

Equations 6.24 to 6.26 show that in the presence of estimation errors in echo RTFs, the residual echo PSD estimate remains independent from the speech component: the multiplicative factor $P(k,i)$ in Equation 6.26 is independent of the near-end speech component. Errors in the computation of the echo RTF do not lead to leakage of the near-end speech in the residual echo PSD estimate. Given the expression $P(k,i)$, we cannot presume that the echo PSD is going to over- or under-estimated. We can simply state from Equation 6.26 that the errors in the echo RTF will simply result in over- or underestimation of the residual echo. Assuming that the loudspeaker and near-end components are not correlated, we can state that the echo RTF estimate does not impact on the near-end speech signal but only the amount of echo suppression achieved by the proposed postfilter.

6.1.2.3 Summary of the proposed residual echo PSD estimate

In this section, we propose an approach to estimate the residual echo PSD in DM mobile terminals. The proposed estimate is expressed as a function of two RTFs which need to be estimated. Later on the impact of these RTF estimates is studied. It results from this analysis that:

- **During echo-only periods:** Considering that the RTF estimation errors both occur at the same time, as will be the case in the real time system, the residual echo PSD becomes:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = P(k,i) \cdot \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) \tag{6.27}$$

$$\text{with} \quad P(k,i) = \frac{|\Theta(k,i) + \Delta\Theta(k,i)|^2 - |\Gamma(k,i)|^2}{|\Theta(k,i) + \Delta\Theta(k,i)|^2 - |\Gamma(k,i) + \Delta\Gamma(k,i)|^2}. \tag{6.28}$$

Note that, no matter what the value of the near-end RTF error $\Delta\Theta(k,i)$, the multiplicative term $P(k,i)$ tends to 1 as $\Delta\Gamma(k,i)$ is small

$$|\Delta\Gamma(k,i)| \to 0 \Rightarrow |P(k,i)| \to 1. \tag{6.29}$$

The accuracy of the proposed PSD estimate is fully defined by the echo RTF accuracy. Errors in the echo RTF estimation are responsible for residual echo PSD over- and under estimation. In practice to compensate for errors in the PSD estimate, the residual echo PSD is weighted with an overestimation factor. The use of an overestimation factor artificially increase the residual echo PSD and permits to ensure complete attenuation of the echo during echo-only periods.

- **During double-talk periods:** Considering both RTF errors at the same time, the residual echo PSD can be written as:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = P(k,i)\cdot\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + \Delta\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) \tag{6.30}$$

$$\text{with} \quad P(k,i) = \frac{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\Gamma(k,i)|^2}{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\Gamma(k,i)+\Delta\Gamma(k,i)|^2} \tag{6.31}$$

$$\Delta\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\Theta(k,i)|^2}{|\Theta(k,i)+\Delta\Theta(k,i)|^2 - |\hat{\Gamma}(k,i)|^2}\cdot\Phi^{s_1s_1}(k,i). \tag{6.32}$$

The accuracy of the residual echo PSD estimate is affected by the accuracy of both the echo and near-end RTFs. Errors in the echo RTFs result in over- and underestimation of the residual echo components $P(k,i)$ while the errors in the near-end RTF introduce some near-end speech components ($\Delta\Phi^{\tilde{d}_1\tilde{d}_1}$) in the residual echo PSD estimate. Assuming that $\Phi^{xs}(k,i) = 0$, we can predict that part of the SA will be due to the inaccuracy of the near-end RTF. Because of the term $\Delta\Phi^{\tilde{d}_1\tilde{d}_1}$, overestimating the residual echo PSD will result in higher SA during DT periods. Therefore the overestimation factor should be chosen carefully.

To avoid inaccurate estimates due to the denominator $||\Gamma|^2 - |\Theta|^2|$, its value should in practice be thresholded so as to be greater or equal to 0.1:

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \frac{|\hat{\Theta}(k,i)|^2\cdot\Phi^{e_1e_1}(k,i) - \Phi^{y_2y_2}(k,i)}{\max(||\hat{\Theta}(k,i)|^2 - |\hat{\Gamma}(k,i)|^2|, 0.1)}. \tag{6.33}$$

The PSDs $\Phi^{e_1e_1}$ and $\Phi^{y_2y_2}$ can be computed directly from the input signal $e_1(n)$ and $y_2(n)$ through autoregressive smoothing.

6.2 Non-linear echo suppression

6.2.1 Problem description

Before introducing the new DM approach to non-linear echo processing we first establish a model of the non-linear echo problem. One suitable model which summarizes the coupling of acoustic sources to two transducers is illustrated in Figure 6.2. The system still has two acoustic sources: near-end speech denoted by $s(n)$ and the non-linear signal $x_{nl}(n)$ emitted by the loudspeaker. The non-linear signal $x_{nl}(n)$ is related to the received loudspeaker or far-end speech signal $x(n)$ through a non-linear transformation which we denote by

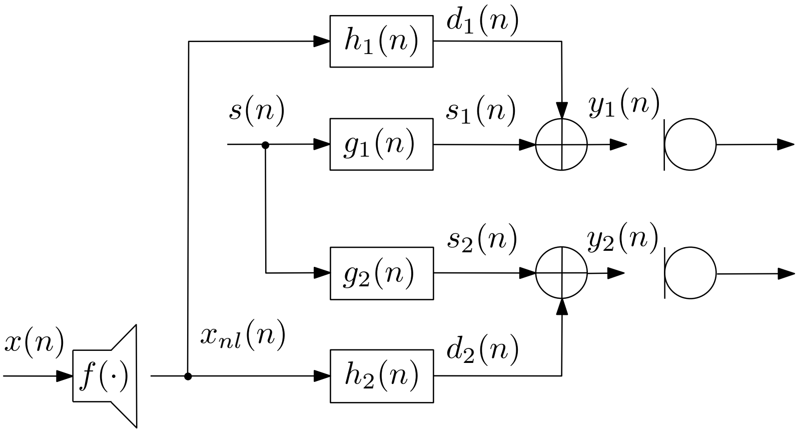


Figure 6.2: Signal model in presence of loudspeaker non-linearities

$f(.)$ such that $x_{nl}(n) = f(x(n))$. The non-linear transformation $f(.)$ is a generic model appropriate to any existing non-linear loudspeaker model such as a clipping model [Stenger and Kellermann, 2000] or a Volterra model [Azpicueta-Ruiz et al., 2011]. For real mobile devices, such modeling accounts for all non-linearities related to the loudspeaker from those due to the cone vibrations to those due to dynamic saturations.

Both near-end speech $s(n)$ and the non-linear signal $x_{nl}(n)$ are recorded by the two microphones, each according to a different acoustic path which we denote similarly as in Chapter 5. The microphone signals are denoted $y_1(n)$ and $y_2(n)$ respectively for primary or secondary microphone signals. The near-end speech signal at the primary and secondary microphone is denoted $s_1(n)$ and $s_2(n)$ and is the result of convolution with acoustic path $g_1(n)$ and $g_2(n)$ respectively. The non-linear loudspeaker $x_{nl}(n)$ also encounters different acoustic paths denoted by $h_1(n)$ and $h_2(n)$ to produce signals $d_1(n)$ and $d_2(n)$ at the primary and secondary microphone respectively. In order words, the microphone signals $y_1(n)$ and $y_2(n)$ are expressed as follows:

$$\begin{aligned} y_j(n) &= s_j(n) + d_j(n) \\ &= g_j(n) * s(n) + h_j(n) * x_{nl}(n) \end{aligned} \tag{6.34}$$

where $j \in \{1, 2\}$.

6.2.2 Limitations of the proposed DM PSD estimate

We want to achieve echo cancellation using the echo processing shown in Figure 6.1. In [Mossi et al., 2010], it is showed that although the AEC is less disturbed by the presence of the non-linearities than by the presence of ambient noise. Because of the correlation between the linear and non-linear components of the echo, the AEC is still able to estimate the acoustic path. The AEC can still be used to estimate the linear part of the echo. The error signal at the output of the AEC can be written as:

$$e_1(n) = y_1(n) - \hat{h}_1(n) * x(n) \tag{6.35}$$

$$= s_1(n) + h_1(n) * x_{nl}(n) - \hat{h}_1(n) * x(n) \tag{6.36}$$

$$= s_1(n) + \tilde{d}_1(n) \tag{6.37}$$

where $\tilde{d}_1(n)$ denotes the residual echo which we consider to be composed of the non-linear echo and of the residual linear echo. In this context, the DM postfilter aims to suppress the non-linear part of the echo as well as the residual linear echo. We consider the same gain rule as in Section 6.1.The required residual echo PSD $\Phi^{\tilde{d}_1\tilde{d}_1}$ is defined, from equations 6.36 and 6.37, as:

$$\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) = |\hat{H}_1|^2 \cdot \Phi^{xx}(k,i) + |H_1|^2 \cdot \Phi^{x_{nl}x_{nl}}(k,i) - 2\,\mathrm{Re}(\hat{H}_1(k,i) \cdot H_1(k,i)^* \cdot \Phi^{xx_{nl}}(k,i)) \tag{6.38}$$

where $\Phi^{x_{nl}x_{nl}}$ is the PSD of $x_{nl}(n)$ and $\Phi^{xx_{nl}}$ is the cross-PSD between $x(n)$ and $x_{nl}(n)$. The first term of the sum $|\hat{H}_1(k,i)|^2 \cdot \Phi^{xx}(k,i)$ can be computed using the impulse response estimate from the AEC $\hat{h}_1(n)$ and the received loudspeaker signal $x(n)$. The two other terms of the sum need to be estimated as $h_1(n)$ and $x_{nl}(n)$ are unknown.

The PSD of the echo signal present at the secondary microphone can be written as:

$$\Phi^{d_2d_2}(k,i) = |H_2(k,i)|^2 \cdot \Phi^{x_{nl}x_{nl}}(k,i). \tag{6.39}$$

In Section 6.1, we used the echo RTF Γ to link $\Phi^{\tilde{d}_1\tilde{d}_1}$ and $\Phi^{d_2d_2}$ and to derive an estimate of $\Phi^{\tilde{d}_1\tilde{d}_1}$. Here we cannot link the two PSDs mainly because of the cross-PSD term $\Phi^{xx_{nl}}$ which is present in Equation 6.38 but not in Equation 6.39. It cannot be computed since $x_{nl}(n)$ is unavailable in the system. To circumvent this problem, we define the residual echo PSD differently. Two options are considered and presented in the following.

6.2.3 Residual echo PSD estimate in presence of loudspeaker nonlinearities

6.2.3.1 Residual echo PSD as a function of the error signal

Based on Equation 6.35, we can define the residual echo PSD $\Phi^{\tilde{d}_1\tilde{d}_1}$ as a function of the error signal PSD:

$$\Phi^{\tilde{d}_1\tilde{d}_1}(k,i) = \Phi^{e_1e_1}(k,i) - \Phi^{s_1s_1}(k,i). \tag{6.40}$$

The definition in Equation 6.40 is valid under the assumption that $x(n)$ and $s(n)$ are decorrelated. We estimate $\Phi^{s_1s_1}$ through a Wiener filter which we apply to the primary microphone signal $y_1(n)$ as follows:

$$\hat{\Phi}^{s_1s_1}(k,i) = \tau \cdot V^2(k,i) \cdot \Phi^{y_1y_1}(k,i) \quad \text{with} \quad V(k,i) = \frac{\psi(k,i)}{1+\psi(k,i)} \tag{6.41}$$

$$\psi(k,i) = \beta \frac{\hat{\Phi}^{s_1s_1}(k-1,i)}{\hat{\Phi}^{d_1d_1}(k-1,i)} + (1-\beta) \cdot \max\left(\frac{\Phi^{y_1y_1}(k,i)}{\hat{\Phi}^{d_1d_1}(k,i)} - 1, 0\right) \tag{6.42}$$

where V is a Wiener filter, τ is a dynamic amplification gain and ψ denotes the SER before AEC. More specifically ψ can be defined as the ratio between $\Phi^{s_1s_1}$ and $\Phi^{d_1d_1}$. ψ is estimated through decision directed approach and therefore its computation only requires an estimate of the echo PSD $\Phi^{d_1d_1}$.

Similarly to Equation 6.39, the primary microphone echo PSD can be defined as

$$\Phi^{d_1d_1}(k,i) = |H_1(k,i)|^2 \cdot \Phi^{x_{nl}x_{nl}}(k,i). \tag{6.43}$$

Assuming once more that $s(n)$ and $x(n)$ are decorrelated, the PSDs of the microphone signals $y_j(n)$ with $j \in \{1,2\}$ can be written as:

$$\begin{aligned}\Phi^{y_jy_j}(k,i) &= \Phi^{s_js_j}(k,i) + \Phi^{d_jd_j}(k,i)\\ &= |G_j|^2 \cdot \Phi^{ss}(k,i) + |H_j|^2 \cdot \Phi^{x_{nl}x_{nl}}(k,i)\end{aligned} \tag{6.44}$$

Let us introduce a slightly different echo RTF to that used previously

$$\Gamma^{'}(k,i) = \frac{H_2(k,i)}{H_1(k,i)}. \tag{6.45}$$

The echo RTF $\Gamma^{'}$ defines the ratio between the frequency responses of the echo paths before the AEC at the contrary of Γ which defines the RTF after the AEC. By using the echo RTF $\Gamma^{'}$ and the near-end RTF Θ, the secondary microphone PSD can be rewritten as:

$$\Phi^{y_2y_2}(k,i) = |\Theta(k,i)|^2 \cdot |G_1(k,i)|^2 \cdot \Phi^{ss}(k,i) + |\Gamma^{'}(k,i)|^2 \cdot |H_1(k,i)|^2 \cdot \Phi^{x_{nl}x_{nl}}(k,i). \tag{6.46}$$

An estimate of $\Phi^{d_1d_1}$ is derived by combining the definitions of $\Phi^{y_1y_1}$ and $\Phi^{y_2y_2}$ (Equations 6.44 and 6.46):

$$\hat{\Phi}^{d_1d_1}(k,i) = \frac{\left||\Theta(k,i)|^2 \cdot \Phi^{y_1y_1}(k,i) - \Phi^{y_2y_2}(k,i)\right|}{\max(||\Theta(k,i)|^2 - |\Gamma|^2|, 0.1)}. \tag{6.47}$$

Once more, the denominator is thresholded to avoid problems due to division by small numbers. $\hat{\Phi}^{d_1d_1}$ computed as in Equation 6.47 is used to compute the SER ψ and later on $\hat{\Phi}^{s_1s_1}$.

The amplification gain τ in Equation 6.41 is used to avoid excessive attenuation of near-end speech during double talk. In our implementation, it is controlled using the normalized PLD described in Section 5.3 (See Equation 5.18). In echo-only periods, τ is set to 1 while during double-talk periods τ is set to 2.

6.2.3.2 Residual echo PSD as a function of the echo picked by the primary microphone

The echo recorded by the primary microphone can be written as the sum of its linear estimate from the AEC and of the residual echo

$$d_1(n) = \hat{d}_1(n) + \tilde{d}_1(n). \tag{6.48}$$

Equivalently, its PSD can be defined as:

$$\Phi^{d_1d_1}(k,i) = \Phi^{\hat{d}_1\hat{d}_1}(k,i) + \Phi^{\tilde{d}_1\tilde{d}_1}(k,i) + 2\,\mathrm{Re}(\Phi^{\hat{d}_1\tilde{d}_1}(k,i)). \tag{6.49}$$

If we consider that the AEC has reached its optimum, based on the orthogonality principle, the cross-PSD term $\Phi^{\hat{d}_1\tilde{d}_1}$ is null [Haykin, 2002]. In other terms, when the AEC reaches its optimum, the echo estimate $\hat{d}_1(n)$ and the residual echo $\tilde{d}_1(n)$ define two orthogonal sub-spaces of the vectorial space generated by $d_1(n)$. In a real time system, the AEC takes some time to reach its optimum, therefore, the orthogonality principle does not apply. But yet, according to Equation 6.48, we can somehow say $\hat{d}_1(n)$ and $\tilde{d}_1(n)$ generate two complementary sub-spaces of the vectorial space generated by $d_1(n)$. Based on this complementarity, the cross-PSD term is therefore negligible.

By neglecting $\Phi^{\hat{d}_1\tilde{d}_1}$, another estimate of the residual echo can be derived as

$$\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}(k,i) = \Phi^{d_1d_1}(k,i) - \Phi^{\hat{d}_1\hat{d}_1}(k,i). \tag{6.50}$$

The required echo PSD $\Phi^{d_1d_1}$ can be estimated as in Equation 6.47. The PSD of the echo estimate from the AEC $\Phi^{\hat{d}_1\hat{d}_1}$ can be computed through any appropriate method as the signal $d_1(n)$ is known in the system. In our case, $\Phi^{\hat{d}_1\hat{d}_1}$ is computed through autoregressive smoothing.

1.	Update of PSDs and cross-PSDs $\Phi^{y_1y_1}$, $\Phi^{y_2y_2}$, $\Phi^{e_1e_1}$, Φ^{xy_1} and Φ^{xy_2}	
2.	Update of RTFs estimate $\hat{\Theta}$ and $\hat{\Gamma}^{'}$	
3.	Compute $\hat{\Phi}^{d_1d_1}$ (Equation 6.47)	
4.	**PSD estimate according to Section 6.2.3.1** a. Update near-end PSD estimate $\hat{\Phi}^{s_1s_1}$ (Equation 6.41) b. Update residual echo PSD $\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}$ (Equation 6.40)	**PSD estimate according to Section 6.2.3.2** a. Update residual echo PSD $\hat{\Phi}^{\tilde{d}_1\tilde{d}_1}$ (Equation 6.50)

Table 6.1: Summary of the residual echo PSD update procedure

6.2.3.3 Estimation of the echo relative transfer function

Both residual echo PSD estimates proposed in Sections 6.2.3.1 and 6.2.3.2 require the estimation of $\Phi^{d_1d_1}$. The proposed estimate $\hat{\Phi}^{d_1d_1}$ according to Equation 6.47 requires the knowledge of the newly defined echo RTF $\Gamma^{'}$.

The echo RTF $\Gamma^{'}$ can still be estimated through the cross-PSDs between the received loudspeaker signal $x(n)$ and $y_1(n)$ or $y_2(n)$ which we can express as follows:

$$\Phi^{xy_j} = H_j \cdot \Phi^{xx_{nl}}. \tag{6.51}$$

From Equation 6.51, an estimate of $\Gamma^{'}$ is derived as:

$$\hat{\Gamma}^{'} = \frac{|\Phi^{xy_2}|}{|\Phi^{xy_1}|}. \tag{6.52}$$

Lastly, we recall that the computation of $\hat{\Phi}^{d_1d_1}$ only requires the modulus of the echo RTF. Accordingly, phase issues corresponding to the cross-PSDs can be safely ignored. Similarly as for $\hat{\Gamma}$ (See Section 5.4.2.2), we compute $\hat{\Gamma^{'}}$ during far-end speech activity periods which we detect using a threshold on the PSD of the loudspeaker signal $x(n)$. The value of $\Gamma^{'}$ is frozen during the periods of far-end speech inactivity.

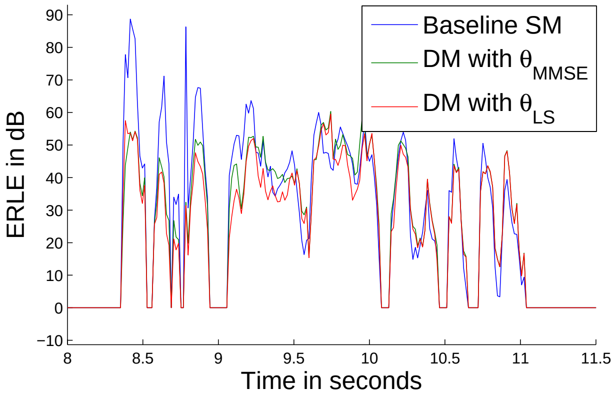
6.2.4 Summary of non-linear echo suppression

In this section, two approaches to residual echo PSD estimation in the presence of loudspeaker non-linearities are presented. They are both based on the estimation of the non-linear echo picked up by the primary microphone which is dependent on RTFs. The proposed residual echo PSD estimation approaches are summarized Table 6.1.

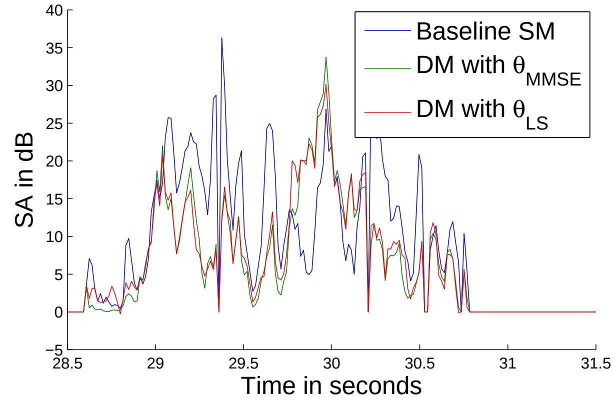
6.3 Experiments

6.3.1 Setup

Our DM echo processing scheme is compared to the SM echo processing scheme used in Chapter 3. The adaptive filter considered is the same as that used in Chapter 5: NLMS adaptive filter [Haykin, 2002] with variable stepsize [Benamar, 1996]. Both DM and SM postfilters use the same gain rule (i.e. Wiener filter with SER estimated through DDA). The DM and SM postfilters differ by the residual echo PSD estimator.



(a) PSD accuracy in echo only periods



(b) PSD accuracy in double-talk periods

Figure 6.3: Objective performance of the DM echo estimate for different RTF estimation methods

The assessment is carried out in two steps. On one hand, the accuracy of the PSD estimate is evaluated using the mean of symmetric segmental logarithmic error (See Chapter 2). On the other hand, echo suppression performance is assessed in terms of ERLE and SA.

6.3.2 Linear echo processing

Experiments involve the same signals as in Chapter 5. The different impulse responses recorded with our mock-up phone are used to generate a test database of speech signals with SERs ranging from -5 to $10dB$ on the primary microphone. Linear and non-linear echo suppression are assessed in Section 6.3.2 and 6.3.3 respectively. The DM residual echo PSD estimate considered in this section is that of Section 6.1.

6.3.2.1 Influence of the near-end RTF estimate

The proposed DM echo estimate depends on the near-end and echo RTFs. In Chapter 5 (see Section 5.4.2) two near-end RTF estimates are presented: $\hat{\Theta}_{MMSE}$ and $\hat{\Theta}_{LS}$. Figure 6.3 shows the influence of the different near-end RTF estimators on the ERLE and

SA. The ERLE and SA for the SM case is shown for reference. We can see from Figure 6.3 (a) that between time $t = 8.3s$ and $t = 9.5s$, the DM systems achieve less echo suppression than the SM system. This is explained by the fact that this portion corresponds to the beginning of the echo periods. This poor behavior is due to the echo RTF which has not yet reached its steady state. Indeed, the echo RTF is computed through the cross-PSDs Φ^{xe_1} and Φ^{xy_2} which are in turn estimated through autoregressive smoothing. At the beginning of the echo period, these PSDs are still in their transient state. As explained in Section 6.1.1, the echo RTF accuracy has a strong impact on the amount of echo suppression.

Figure 6.3 also shows no matter what near-end RTF estimate is used, the ERLE and SA curves for the DM postfilter are very similar. In Section 6.1.1, we showed that SA during DT is partly due to errors in the near-end RTF. Although, the LS estimate $\hat{\Theta}_{LS}$ is shown in [Shalvi and Weinstein, 1996] to be unbiased and therefore a more mathematically accurate estimate compared to the MMSE estimate, it does not significantly improve the performance of our DM PSD estimate. Given the significant computational complexity and memory requirements needed by $\hat{\Theta}_{LS}$, only $\hat{\Theta}_{MMSE}$ is used in further experiments.

6.3.2.2 Residual echo PSD accuracy

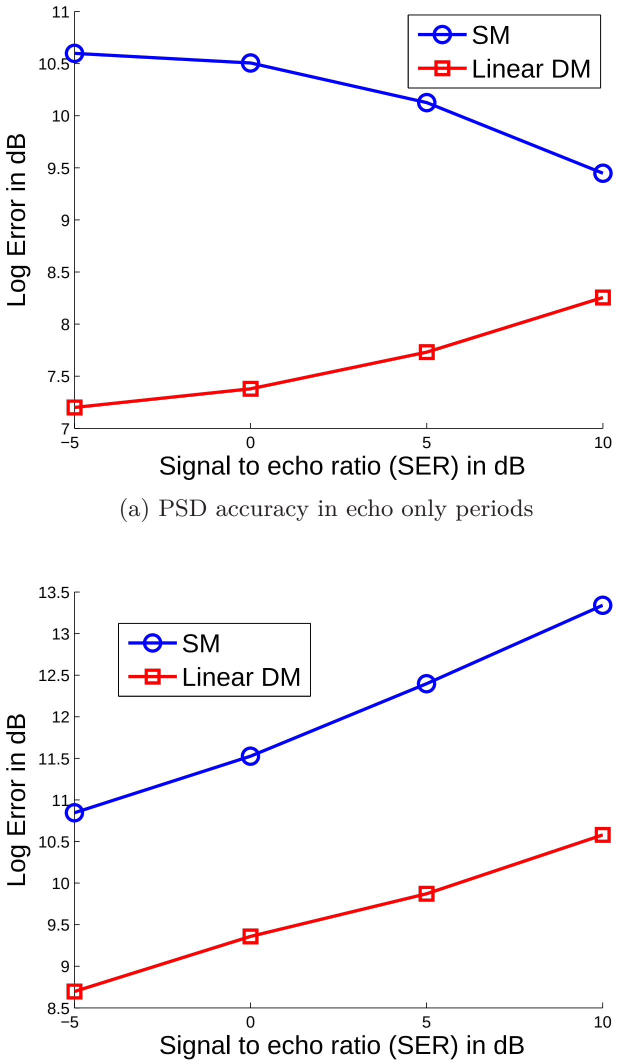
Figure 6.4 illustrates the error in the residual echo PSD estimator during echo-only and double-talk periods. For both echo-only periods and DT periods, the DM estimator outperforms the SM estimator. The curves also show that the error for the DM estimate increases with the SER. For the echo-only periods, this increase is very small (less than 1dB for the whole range of SERs considered). In our opinion, this increase is due to the threshold of the denominator of the DM estimate (Equation 6.33). For the DT periods, the increase is more emphasized and can be explained by the presence of near-end speech, which disturbs the PSD estimate. Moreover, a high SER implies high levels of near-end speech compared to echo (and therefore residual echo) and thus greater disturbance of the residual echo estimators. The loss of performance of the DM estimate can be explained by the fact, that during DT, the presence of near-end disturbs the estimate of the echo RTF as the cross-PSDs used for its computation contain a component dependent on the near-end speech signal. As the SER increases, the cross-PSD component due to the near-end speech signal increases and so-doing leads to a less accurate estimate of the echo RTF.

It is also of interest to note that there is an increase in the DM PSD error between echo-only and DT periods. This corroborates the analysis in Section 6.1.1 which shows that there are some near-end speech components in the residual echo estimate during DT.

6.3.2.3 Echo suppression

Figure 6.5 shows the ERLE and SA curves for echo-only and DT periods respectively. The ERLE curves in Figure 6.5 (a) show that the DM echo postfilter achieves more echo suppression than the SM postfilter. This is a direct consequence of PSD estimate accuracy during echo-only periods. The SA curves show that for the DM case, attenuation of the near-end speech increases with the SER while for the SM case the attenuation decreases. Such increases in the SA are undesirable as it means half-duplex situations. The DM postfilter nevertheless introduces less attenuation (up to 5dB) compared to the SM postfilter.

Informal listening tests show that the DM postfilter yields slightly better near-end speech intelligibility compared to the SM postfilter during double-talk periods. The high



(a) PSD accuracy in echo only periods

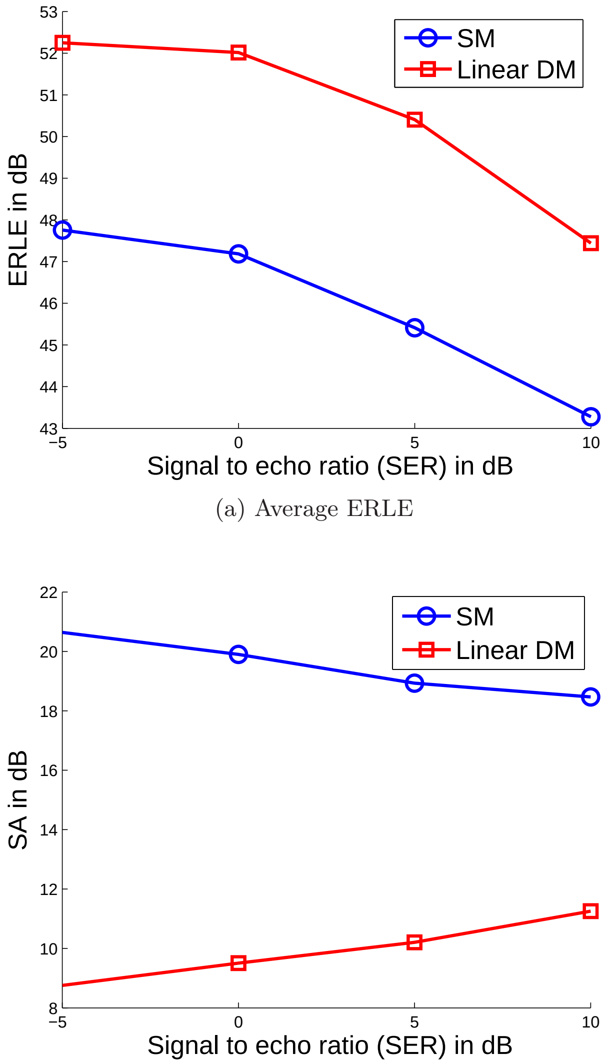
(b) PSD accuracy in double-talk periods

Figure 6.4: Echo PSD estimation error

SA introduced by the SM postfilter is perceptible and sometimes leads to complete suppression of the speech.

6.3.2.4 Assessment with data from real device

Figure 6.6 shows the ERLE and SA curves measured with data from a real device. The SA is measured in a similar way as in Section 5.5.5. Only the ranking of the methods compared and the SA difference between them has a meaning. We see that the SM system achieves the most echo suppression: this observation is confirmed by the spectrograms in Figure 6.7. Nevertheless, it is of interest to note that for both methods, the echo component is completely suppressed during echo-only periods. Bearing in mind, that in a real time system, the postfiltering gains are always floored such that its output level matches that of the ambient noise or comfort noise, we can state that the Linear DM method achieves enough echo suppression.



(a) Average ERLE

(b) Average SA during double-talk periods

Figure 6.5: Echo suppression performance

The spectrograms in Figure 6.7 and the SA curve in Figure 6.6 (b) show the clear superiority of the Linear DM method during double-talk periods. We see from the spectrograms that a lot of the near-end speech components are lost with the SM system.

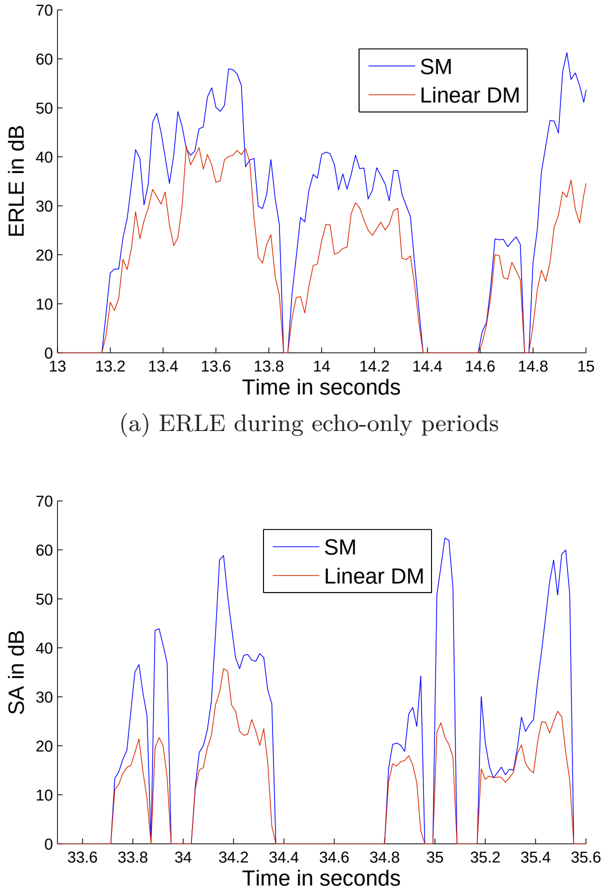
6.3.3 Non-linear echo suppression performance

Non-linear loudspeaker signals are generated according to a Volterra model as follows:

$$x_{nl}(n) = x(n) + a \cdot x^2(n) + b \cdot x^3(n) \text{ with } a = b = 1. \tag{6.53}$$

Our database of impulse responses is used to generate a database of signals with non-linear echo. As before, the average SER at the primary microphone ranges from -5 to 10dB. The resulting database has an average linear-to-non-linear echo ratio ($\sigma_x^2/\sigma_{x-x_{nl}}^2$) of 15dB.

In the following, the proposed residual echo estimates are denoted $NLDM-1$ and $NLDM-2$ for the estimation methods presented in Sections 6.2.3.1 and 6.2.3.2 respec-



(a) ERLE during echo-only periods

(b) SA for a double-talk periods

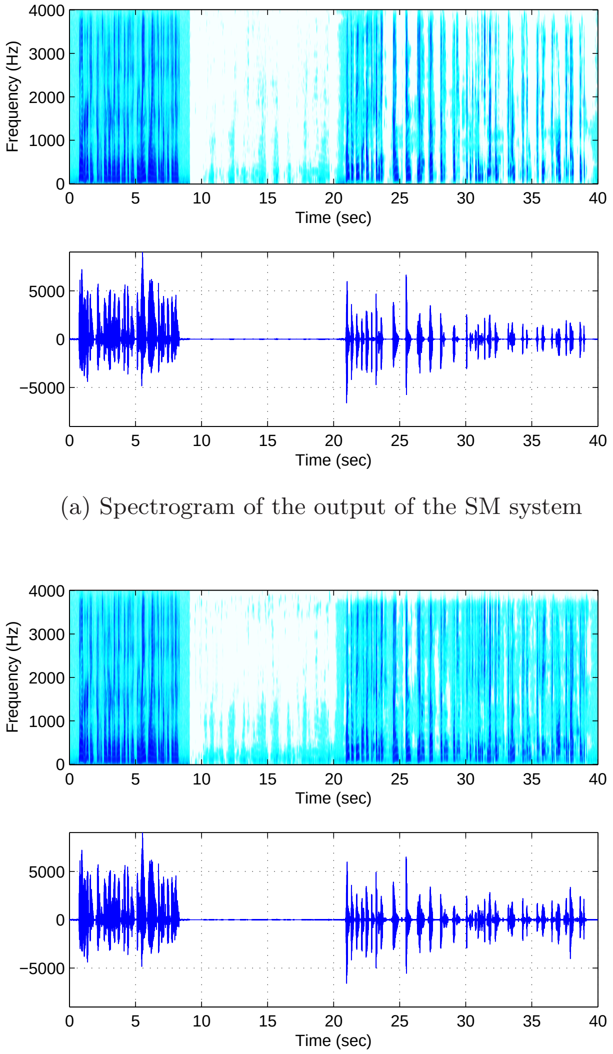
Figure 6.6: Echo suppression performance

tively. Both methods are compared to the SM echo PSD estimate. The linear DM estimate assessed in the previous section is also considered.

6.3.3.1 PSD accuracy

The error in the PSD estimates obtained for each system is illustrated in Figure 6.8 (a) and 6.8 (b) for echo-only and double-talk periods separately. The proposed non-linear echo PSD estimates are showed to outperform the linear estimation methods for both echo-only and double-talk periods. The superiority of the non-linear estimates is even more emphasized in echo-only periods: there is a gap of at least 5dB. The NL-DM-1 system achieves the best performance in echo-only periods whereas in periods of DT, the NL DM-2 is the best. Although the gap between the SM and the linear DM system is lower than that observed in the linear echo case, their relative ranking remains the same.

For all methods considered, the PSD error in DT periods is higher than that in echo-only periods. For each method, we also observe that during DT periods, there is a clear increase in the PSD error with the SER. As for the linear case, the increase in error is due to the presence of the near-end speech signal which leads to relatively poor estimates of the echo RTF. The echo RTF is computed with the cross-PSDs. The more the SER increases, the more the cross-PSD terms are erroneous.



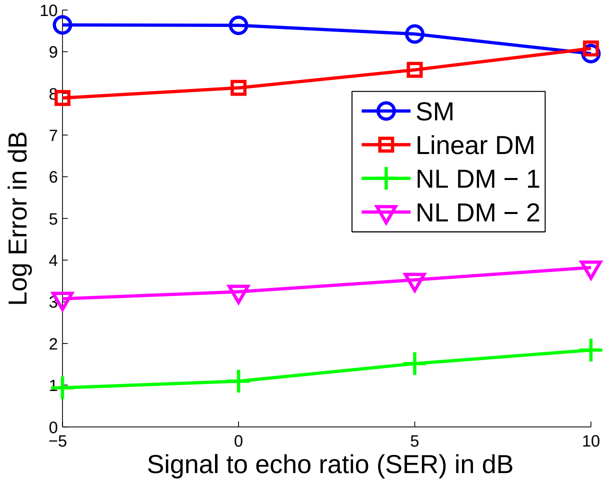
(a) Spectrogram of the output of the SM system

(b) Spectrogram of the output of the DM system

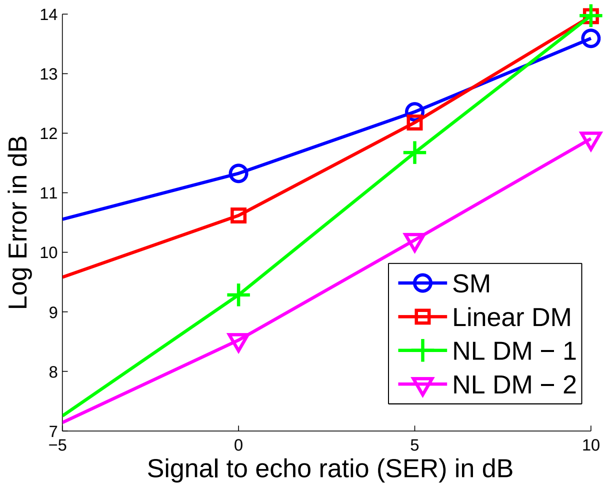
Figure 6.7: Spectrograms of processed signals. Signals are composed of near-end speech only (0 to 9s), echo-only periods (9 to 20s) and double-talk (20 to 40s)

6.3.3.2 Echo suppression

Figure 6.9 (a) shows the echo suppression performance of the 4 different schemes considered. We see that the NL DM-2 system achieves the most echo suppression. The linear DM system seems very disturbed by the presence of non-linearity at low SER. For SER of -5dB, the linear DM system only achieves about 37dB echo attenuation. This poor performance comes from the fact that the lower the SER, the higher the echo level and the more residual echo we have at the output of the AEC. In addition to the limitations of the Linear DM PSD estimate presented in Section 6.2.2, the echo RTF estimate is disturbed by the presence of non-linearities which explains the poor performance. Nevertheless, we observe that the amount of echo suppression achieved by the Linear DM approach significantly increases with the SER. At high SERs (i.e. 10dB), it achieves as much echo attenuation as in the case of linear echo. The NL DM-1 method achieves slightly more echo suppression



(a) PSD accuracy in echo only periods

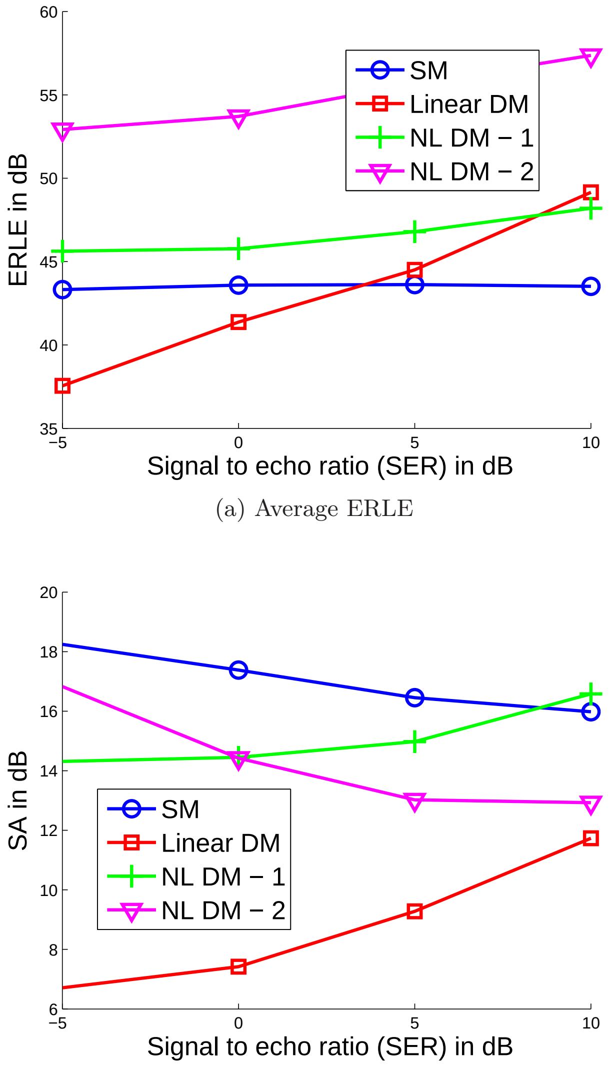


(b) PSD accuracy in double-talk periods

Figure 6.8: Echo PSD estimation error

than the SM system.

SA performance is reported in Figure 6.9 (b). We see that the SM method introduces the most SA while the linear DM system leads to the least SA. The good SA performance of the Linear DM system needs to be put into perspective as this system achieves the least echo suppression. One can increase the amount of echo suppression achieved by this method by using an overestimation factor but this will automatically result in higher SA during DT periods. The proposed non-linear method NL DM-2 leads to slightly less SA than the SM system. By comparing the ERLE and SA achieved by the NL DM-2 method, we can state that this method leads to good non-linear echo suppression. The level of SA caused by the NL DM-1 system is also lower than that of the SM system but increase as the SER increases. This may be due to the way we estimate the non-linear echo PSD in this case, through spectral subtraction based on an estimate of the near-end PSD measured before the AEC.



(a) Average ERLE

(b) Average SA during double-talk periods

Figure 6.9: Echo suppression performance

6.4 Conclusion

This chapter focuses on echo postfiltering for DM devices. Similarly as in Chapter 5, residual echo suppression only takes place on the error signal coming from the AEC. The DM microphone signals are used to estimate the PSD of the residual echo which we used to compute the residual echo attenuation gain. The logic used to estimate the residual echo PSD is extended to non-linear echo suppression. Our investigations of the non-linear echo suppression lead to two new approaches to estimate the non-linear residual echo PSD.

Assessment shows that the proposed linear PSD estimate leads to more echo suppression than the SM method and a significant reduction of the amount of distortion during DT periods. The proposed linear PSD estimate collapses in presence of non-linearities. The non-linear estimators proposed yield good echo suppression, especially that of Section 6.2.3.2 (NL DM-2).

Chapter 7

Conclusion and perspectives

Speech quality in mobile devices is degraded by acoustic echo and ambient noise. The annoyance due to both disturbances increases as their level increases. Speech processing algorithms need to be implemented within mobile phones to maintain good speech quality. While both are addressed at some point in this thesis, we focus on approaches to improve echo control for single- and dual-microphone devices.

7.1 Single microphone echo control

Existing approaches to single microphone echo control are based on adaptive filtering followed by echo postfiltering. Adaptive echo cancellation aims to estimate the echo signal recorded by the microphone while echo postfiltering consists in attenuating the residual echo in the frequency or subband domain. Similarly noise reduction approaches aim to attenuate the noise in the frequency or subband domain.

A simple approach to reduce the computational complexity of a speech enhancement scheme consists in combining echo postfiltering and noise reduction. In our case, we focus mainly on efficient filtering methods (i.e. perturbation suppression). Subband domain filtering suffers from frequency domain aliasing which can be avoided by performing the filtering in the time domain. Our assessment shows that both subband and time domain filtering methods are equivalent in terms of perturbation attenuation. Perceived distortion is nevertheless different. Filtering in the time domain tends to introduce crackling noise during DT periods whereas subband filtering results in musical noise.

Echo postfiltering and noise reduction can also be combined in the frequency domain. Filtering in the frequency domain sometimes corresponds to circular convolution which suffers from time domain aliasing. We show how this aliasing can be avoided with appropriate zero-padding. This zero-padding, also referred to as the linear convolution, requires additional DFT and is thus very computational demanding. We proposed an alternative scalable filtering method - the scalability of the proposed method can be exploited to reduce the computational complexity. Comparison of these frequency domain related filtering methods show that they are similar in terms of perturbation attenuation and near-end speech distortion. The proposed systems is therefore a good alternative to circular convolution in the frequency domain.

Another contribution reported in this thesis refers to synchronized echo control system. Our approach to synchronize AEC and echo postfiltering is based on the link between the AEC system mismatch and its power spectrum. Assessments show the proposed synchronization method results in significant reduction in convergence time of the AEC and to better echo suppression.

7.2 Dual-microphone echo control

The focus in Part II of this thesis relates to echo control for DM mobile devices. An analysis of the echo problem based on recordings with DM devices is reported. It results from this study that by placing the transducers in a certain configuration, significant level differences an be observed between the microphone signals depending on whether we are in echo-only, near-end speech only or DT periods.

Unlike existing multi-microphone echo control systems, we choose to use an adaptive filter followed by echo postfiltering. In Chapter 5, we show how the level difference observed with our recordings can be exploited for DT detection. A new DM echo postfiltering gain rule is also proposed. Assessment show the proposed DTD can be used efficiently to improve any existing SM echo control algorithm. We also show that the use of the proposed DTD in conjunction with the new gain rule yields good echo attenuation as well as low distortion of the useful near-end speech signal. We show in Chapter 5 that the performance of the proposed DTD and gain rule significantly increase as the level difference between the microphones signals also increases. The performance of methods based on level difference lead to recommendations regarding hardware configurations of the transducers on the devices. The primary microphone should be placed as far as possible from the loudspeaker while the secondary microphone should be placed as close as possible to guarantee maximum level difference between microphone signals.

The methods presented in Chapter 5 are bounded to function for certain transducers configuration. In Chapter 6 we proposed a more general approach to DM echo postfiltering. This method exploits the correlation between the microphone signals and can be extended to non-linear echo suppression. We show through experiments that this method outperforms the SM approach in terms of echo attenuation and speech distortion.

7.3 Perspectives

While SM echo control solutions have largely been investigated in the literature, our contributions show that there is still room for speech quality improvement. The synchronization approach presented in this thesis is based on a NLMS adaptive filter. The synchronization approach could be interest to improve convergence speed of the APA.

The proposed DM approaches to echo postfiltering all use estimates of the echo and near-end RTFs. Our analyzes show that the accuracy of these RTFs highly impact on the echo suppression performance of the proposed methods. RTF estimates can be found in the literature: they are used within multi-microphones speech enhancement algorithms based on beamforming [Gannot et al., 2001, Reuven et al., 2007a]. One of the weakness related to RTF is that it is difficult to its target or ideal value since it is a ratio of impulse responses. Since, it is difficult to define the target value of a RTF, its estimates cannot be evaluated separately but are evaluated as part of a speech enhancement module as we did in Chapters 5 and 6. Methods to improve the proposed DM postfilters should carry on more accurate estimation approaches of the RTF.

The proposed DM methods presented in this thesis do not account for the presence of ambient noise. Future investigations regarding these methods should assess their performance and robustness in presence of noise. Moreover the proposed DM methods are compared to SM methods. Further validation of the proposed methods should include comparison of their performance with existing beamforming approaches.

It is well known that hard decision systems such as the proposed DTD suffers from false positive (i.e. DT declared whereas there is no DT) and false negative (i.e. DT not declared whereas it is DT) errors that are due to thresholding. The proposed measurement of the normalized PLD could also be directly used to control the AEC and/or echo postfiltering through soft decision methods. In such case the stepsize or echo suppression gain could be computed or postprocessed based on the value of the PLD. This soft-decision methods make even more sense given the fact our analysis of the normalized PLD is a function of the signal-to-echo ratio (see Section 5.3).

Annulation d’écho acoustique pour terminaux mobiles à un ou deux microphones

Contents

A.1 Introduction

En physique, l'écho se définit comme étant le signal résultant de la réflexion d'une onde dans son environnement. Ce phénomène peut se produire dans les télécommunications et on parle alors d'écho acoustique. Dans une conversation téléphonique, l'écho acoustique est dû au couplage entre le haut-parleur et le microphone [Hänsler and Schmidt, 2004]. En conséquence, le microphone du téléphone capte tout aussi bien la voix du locuteur local que le signal d'écho généré par le haut-parleur. De plus, lorsque la communication a lieu dans un environnement bruité, le microphone va également capter une partie du bruit ambiant [Boll, 1979]. Si aucun traitement n'est effectué, le signal microphonique transmis au locuteur lointain contient le signal de parole utile mais aussi le signal d'écho et de bruit ambiant. Ainsi le locuteur lointain entendra une version retardée de sa propre voix. La perception de l'écho et du bruit est particulièrement gênante pour le locuteur lointain. La gêne occasionnée par l'écho est d'autant plus forte que le retard introduit par la chaîne de transmission est important et que le niveau de l'écho est élevé.

Afin de garantir une qualité de la parole acceptable , des algorithmes de traitement de la parole doivent être mis en œuvre dans les terminaux mobiles [Degry and Beaugeant, 2008, Hänsler and Schmidt, 2004]. Les techniques d'annulation d'écho acoustique ont été largement étudiées dans la littérature. Les techniques les plus communément employées consistent d'une annulation d'écho adaptative suivi d'un post-filtre [Benesty et al., 2007, Hänsler and Schmidt, 2004, Martin, 1995]. L'annulation d'écho adaptive utilise un filtre adaptif afin de produire une estimation du signal d'écho capté par le microphone. Cette estimation n'étant pas parfaite, il est nécessaire d'utiliser un post-filtre afin de rendre l'écho inaudible [Benesty et al., 2007, Hänsler and Schmidt, 2004].

Les techniques de post-filtrage les plus couramment employées consistent à atténuer le signal d'erreur provenant de l'annulation d'écho adaptative. Afin d'obtenir de bonnes performances en périodes de double-parole, cette atténuation doit être calculé dans le domaine fréquentiel ou sous-bandes. Malgré tout, les performances de la d'annulation d'écho telles que décrites ici restent encore limitées : il y'a toujours un compromis à faire entre la suppression d'écho pendant les périodes d'écho seul et la quantité de distorsions introduites sur le signal utile en période de double-parole.

La réduction de bruit quant à elle est basée sur l'hypothèse selon laquelle le bruit est une perturbation additive qui a des propriétés spectrales différentes de celle de la parole. Les algorithmes les plus courant consiste en appliquer une atténuation au signal bruite dans le domaine fréquentiel ou en sous-bandes [Gustafsson et al., 2001, Hänsler and Schmidt, 2004, Martin, 2001].

Ces quelques exemples montrent que la qualité vocale d'une communication téléphonique est assurée par un nombre important d'algorithmes de traitement du signal. Historiquement, ces algorithmes ont été développés indépendamment :

- Les différents algorithmes utilisés ne tiennent pas compte des synergies qui peuvent exister entre les différents modules. La réduction de bruit et le post-filtrage de l'écho ont des fonctionnements similaires, ils pourraient être couples afin de réduire la complexité de calcul qui est un aspect déterminant pour les téléphones portables [Oppenheim and Schafer, 1999] .

- De part leur mobilité, les téléphones portables peuvent être utilisés dans des conditions très adverses. En mode mains-libres, le signal d'écho est très fort. Le niveau de bruit devient également très élevé lorsque le téléphone est utilisé en environnement bruité. Il convient donc d'employer des techniques qui pourront efficacement supprimer les perturbations (écho et bruit) tout en préservant le signal utile. Les solutions existantes ne permettent pas d'obtenir un tel compromis [Enzner and Vary, 2003, 2006, Steinert et al., 2007].

- On voit de plus en plus apparaître sur le marché des téléphones portables équipés de deux microphones. Or très peu de solutions de traitement de la parole exploitent cette architecture bi-microphones. Avec une telle architecture, l'on va pouvoir exploiter des propriétés (telles la coherence des signau microphoniques) auxquels l'on ne peut avoir accès dans des systèmes à un microphone [Gannot et al., 2001, Jeub et al., 2011, Kellermann, 1997, Reuven et al., 2007a].

Dans cette thèse, nous étudions l'ensemble de la chaîne de traitement de la parole. Notre objectif est de proposer une architecture ayant une complexité de calcul adapté aux téléphones portables. Cette nouvelle architecture doit également permettre d'améliorer la qualité vocale. Tout d'abord, nous considérons l'annulation d'écho pour terminal à une microphone et proposons une nouvelle architecture qui tient compte des interactions entre l'annulation d'écho adaptative et le post-filtrage écho d'une part. D'autre part nous étudions aussi les méthodes optimales de combinaison de la réduction de bruit et du post-filtrage de l'écho. Deuxièmement, nous proposons des approches pour améliorer l'annulation d'écho sur basées sur des architectures a deux-microphone.

Ce résumé est organisé comme suit. La section A.2 présente plus en détail les problématiques liées aux phénomènes de l'écho et du bruit ambiant. L'état de l'art en matière de traitement de la parole pour la téléphonie mobile est présenté en section A.3. Les sections A.4 et A.5 résument nos contributions concernant le traitement de l'écho pour terminaux à un microphone et terminaux à deux microphones respectivement.

A.2 La prise de son dans la téléphonie mobile

En plus de leur fonction de base (i.e. : d'émettre ou recevoir des appels), les téléphones portables offrent aujourd'hui une variété de service à leur utilisateurs. De ce fait, la plupart des téléphones sont donc équipés d'autres composants électroniques tels que l'appareil photo numérique ou GPS. En conséquence le positionnement du haut-parleur (HP) et du microphone sur le terminal n'est pas toujours optimal. Ce qui va avoir un impact sur la qualité vocale. Mais l'élément le plus déterminant pour la qualité vocale d'une communication téléphonique reste l'environnement acoustique dans lequel a lieu la communication. Dans cette partie nous décrivons les problèmes de l'écho acoustique et du bruit ambiant.

A.2.1 Qu'est-ce que l'écho acoustique?

Le son émis par le locuteur lointain est transmis via le réseau vers le terminal du locuteur local qui le diffuse par le haut-parleur (HP) dont la puissance est suffisante pour que la voix du locuteur distant soit parfaitement audible et compréhensible pour le locuteur local. La voix du locuteur local est quant à elle captée par le(s) microphone(s) puis transmise au locuteur lointain. Dans certains cas, une partie du signal émis par le HP est captée par le(s) microphone(s) et est renvoyée au locuteur distant qui entend ainsi sa propre voix : c'est ce que l'on appelle l'écho acoustique. D'un point de vue acoustique, le son émis par le HP se propage dans le milieu environnant, et il soit directement capté par le microphone

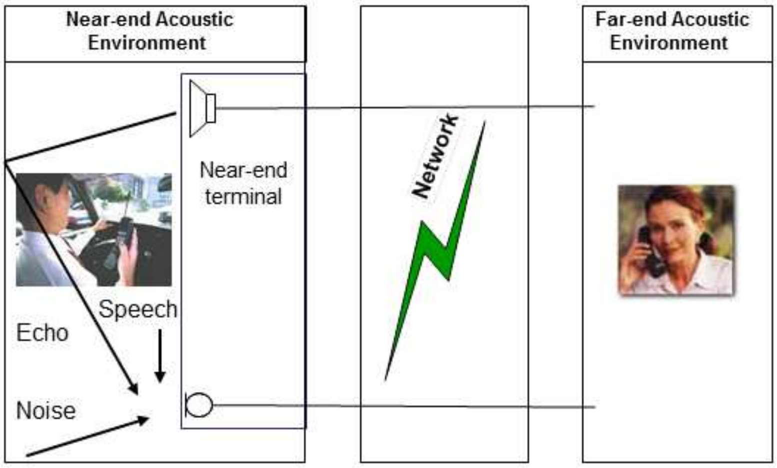


Figure 1: Illustration du problème de l'écho acoustique et du bruit ambiant.

(trajet direct), soit après une ou plusieurs réflexions sur les parois environnantes. La Figure 1 schématise le phénomène d'écho.

A.2.1.1 L'écho linéaire

Le couplage entre le HP et le microphone du terminal est généralement modélisé par un filtre linéaire à réponse impulsionnelle finie, la longueur du filtre étant caractéristique du milieu environnant

$$d(n) = h(n) \star x(n) \tag{1}$$

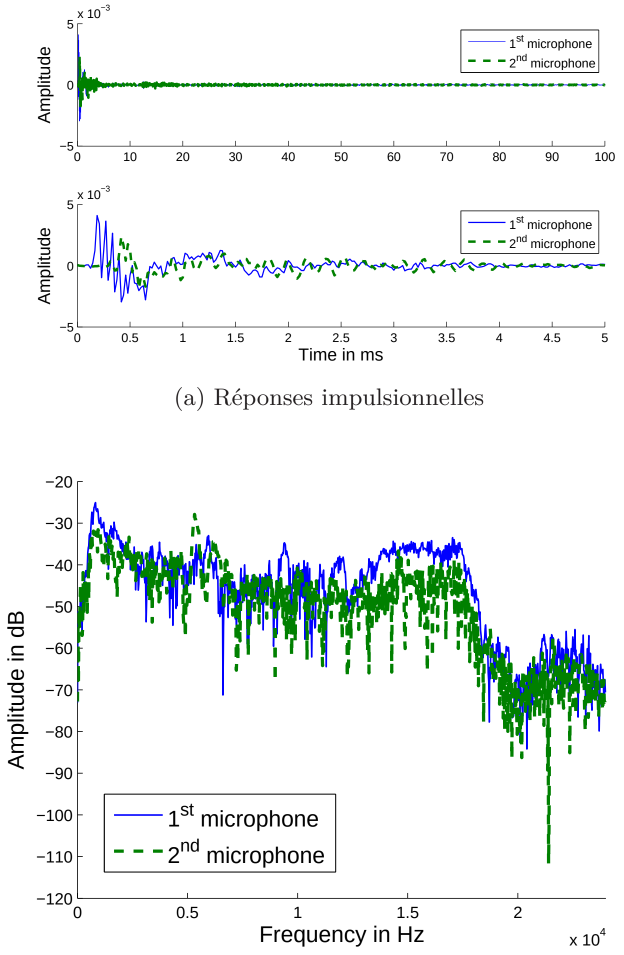
Où $x(n)$ représente le signal HP, $d(n)$ représente signal d'écho capté par le microphone et $h(n)$ représente le canal acoustique.

La Figure 2 montre des exemples de canal acoustique. Les mesures illustrées ici ont été obtenues à partir d'un prototype de téléphone portable que nous avons équipé de deux microphones :

- Le premier microphone étant plus proche du HP
- Le deuxième microphone étant un peu plus éloigné HP que le premier microphone.

La réponse impulsionnelle du canal acoustique montre qu'il est formé du direct chemin entre le haut-parleur et le microphone d'une part et des chemins indirects d'autre part. Nous constatons que le retard principal (premier pic) n'est pas le même pour chaque microphone. Pour chaque réponse impulsionnelle, le principal retard est lié à trajet direct (c'est à dire à la distance) entre le haut-parleur et le microphone considéré [Kuttruff, 2000]. Plus le microphone du HP est proche du microphone, plus court sera ce trajet direct. Notons également que l'amplitude de ce pic est inversement proportionnelle la distance séparant le HP du microphone.

Les pics qui suivent le principal sont dus aux réflexions du son provenant du haut-parleur dans le milieu environnant. Nous pouvons voir sur la Figure 2 (a) que les réflexions sont différentes pour chaque microphone. Les microphones sont placés à des positions différentes sur le terminal et ne captent pas les mêmes réflexions. Le son du haut-parleur se propage dans toutes les directions créant ainsi un nombre infini d'ondes réfléchis qui



(a) Réponses impulsionnelles

(b) Réponses fréquentielles

Figure 2: Exemples de canal acoustique

seront capt
’es par les microphones.

Les réponses fréquentielles des canaux acoustiques mesurés sont également montrées dans la Figure 2 (b). Elles nous montrent l’impact du canal acoustiques sur le signal de haut-parleur : toutes les fréquences ne sont pas également atténuées.

A.2.1.2 L’écho non-linéaire

En plus des phénomènes linéaires décrits précédemment, l’écho peut être créé par des phénomènes non-linéaires : saturations au niveau de l’amplificateur du HP, distorsion du HP lorsqu’il est soumis à des voltages trop importants, couplage solide entre le HP et le microphone [Birkett and Goubran, 1995a, Gao and Snelgrove, 1991, Guerin, 2002]. La Figure 3 montre une mesure des distorsions harmoniques qui peuvent être introduites par le HP. On observe notamment que la réponse du HP n’est pas toujours linéaire pour les signaux de basses fréquences :

- Plus l’amplitude du signal est importante, plus les distorsions seront importantes

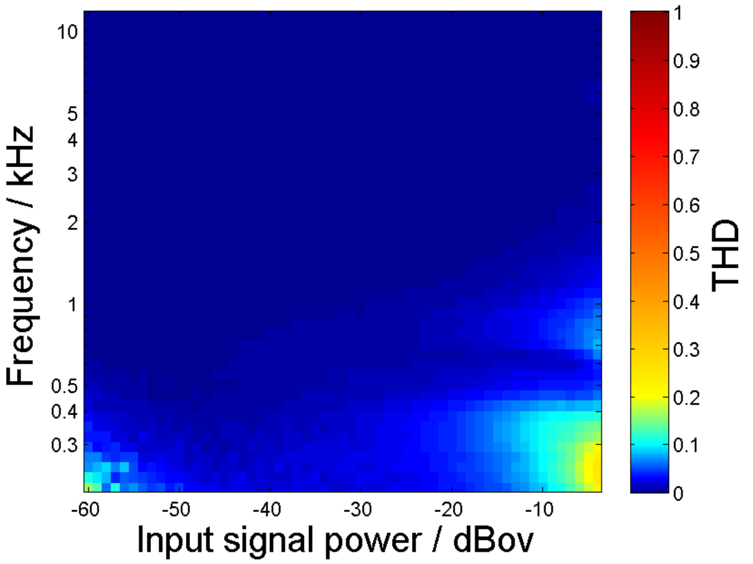


Figure 3: Exemple de mesure de THD (Total Harmonic Distortion) d'un HP

- Les distorsions du signal de faible amplitude sont spécifiques au haut-parleur utilisé et sont liées à ses limites de quantification.

A.2.2 Qu'est ce que le bruit ?

Lors de l'utilisation d'un terminal mobile dans un environnement bruité, une partie du bruit ambiant est également capté par le microphone et transmis au locuteur lointain. Cela peut être très gênant pour le locuteur lointain parce que la parole utile serait partiellement masquée par le bruit ambiant. La gêne due au bruit augmente avec le niveau de bruit. Les exemples les plus communs de source de bruit sont le bruit de voiture, bruit de bureau, bruit de café.

Nous avons expliqué comment le bruit ambiant et l'écho acoustique peuvent dégrader la qualité de la parole dans les communications téléphoniques. Dans la partie suivante, nous allons faire une brève présentation des algorithmes de l'état de l'art qui peuvent être utilisés pour améliorer la qualité de la parole.

A.3 Solutions existantes

A.3.1 L'annulation d'écho acoustique

L'annulation d'écho est illustrée en Figure 4. Elle se fait généralement en deux étapes : l'annulation d'écho adaptative suivi d'un post-filtre. L'annulation d'écho adaptive s'attache à fournir une estimation du signal d'écho capté par le microphone. Il subsiste néanmoins de l'écho résiduel en sortie du module d'annulation d'écho adaptative. Le post-filtre a pour objectif de rendre l'écho résiduel inaudible pour le locuteur lointain.

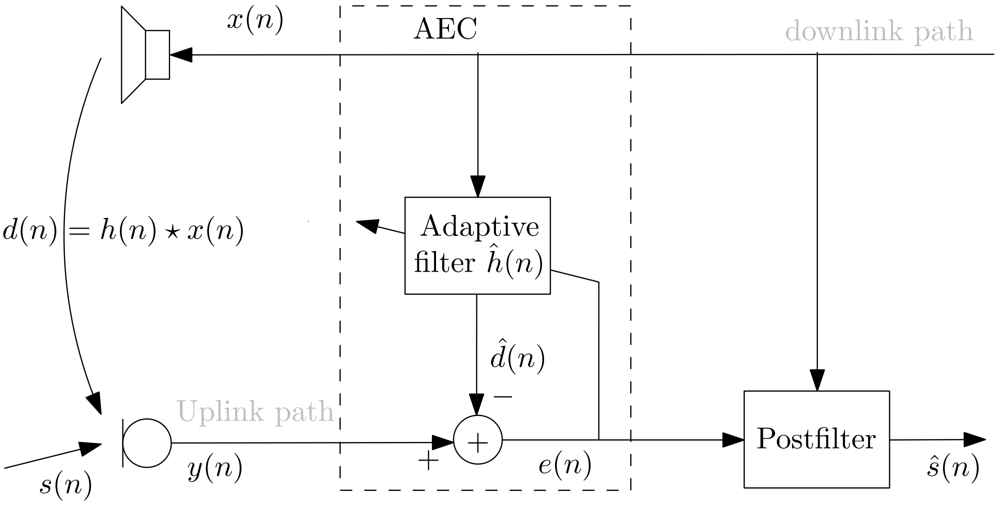


Figure 4: Schéma de l'annulation d'écho

A.3.1.1 L'annulation d'écho adaptative

L'écho acoustique résulte du couplage entre le HP et le microphone du terminal. Ce couplage est modélisé par un filtre à réponse impulsionnelle finie. L'annulation d'écho adaptive utilise un filtre adaptatif pour fournir une estimation $\hat{h}(n)$ de ce couplage. Une estimée de l'écho $\hat{d}(n)$ est ensuite obtenu en convoluant l'estimée du canal acoustique avec le signal de HP [Haykin, 2002]:

$$\hat{d}(n) = (\hat{h} \star x)(n) = \hat{\mathbf{h}}^T(n) \cdot \mathbf{x}(n). \tag{2}$$

où $\mathbf{x}(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-L+1) \end{bmatrix}^T$ est un vecteur contenant les échantillons du signal HP et L est la longueur du filtre adaptatif. L'avantage de l'annulation d'écho adaptative est qu'elle n'affecte en rien la parole utile qui reste parfaitement audible, garantissant ainsi un full-duplex parfait. Le filtre d'estimation du canal acoustique $\hat{h}(n)$ est mis à jour à chaque échantillon par une contre-réaction sur l'erreur d'estimation proportionnellement à un gain d'adaptation $C(n)$:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \hat{\mathbf{C}}(n) \cdot e(n) \quad \text{with} \quad e(n) = y(n) - \hat{d}(n). \tag{3}$$

Le gain d'adaptation C(n) dépend de l'algorithme d'adaptation :

- LMS (Least Mean Square) pour lequel $C(n)$ s'exprime comme suit :

  $$C(n) = \mu \cdot \mathbf{x}(n). \tag{4}$$

  où μ est le pas d'adaptation.

- NLMS (Normalized LMS) pour lequel $C(n)$ s'exprime comme suit :

  $$C(n) = \frac{\mu}{\mathbf{x}^T(n) \cdot \mathbf{x}(n)} \cdot \mathbf{x}(n). \tag{5}$$

  En comparaison au LMS, le NLMS présente l'avantage d'être indépendant du signal HP. Cette indépendance lui confère une convergence qui ne dépend plus que du pas d'adaptation μ.

- APA (Affine Projection Algorithm) pour lequel $C(n)$ s'exprime comme suit :

  $$C(n) = \mu\mathbf{X}(n)\big(\mathbf{X}(n)\mathbf{X}(n)\big)^{-1} \tag{6}$$

  Où $\mathbf{X}(n) = \begin{bmatrix}\mathbf{x}(n) & \mathbf{x}(n-1) & \cdots & \mathbf{x}(n-N+1)\end{bmatrix}$ est une matrice contenant les échantillons du signal de HP et N est l'ordre de l'APA. N représente aussi la fenêtre d'observation. Le principal avantage de l'APA est sa rapidité de convergence en comparaison du LMS ou NLMS mais sa complexité de calcul reste très importante pour un téléphone portable.

La liste ci-dessus n'est pas exhaustive, il existe une multitude d'algorithmes adaptatifs [Haykin, 2002]. Néanmoins pour toutes ces techniques de filtrage adaptatif, l'estimée du canal doit avoir plusieurs centaines de points afin de fournir une estimation fiable. Ainsi, plus le filtre adaptif sera long, plus l'estimation du canal sera bonne. En contrepartie plus un filtre adaptatif est long, plus son temps de convergence (vers sa réponse optimale) sera longue et plus sa complexité calculatoire sera longue [Huo et al., 2001, Paleologu and Benesty, 2012, Sommen and Jayasinghe, 1988].

Comportement de l'annulation d'écho adaptative Le temps de convergence se définit comme le temps que met le filtre adaptatif pour aller de sa réponse initiale (généralement nulle puisque l'on n'a pas d'à-priori sur le canal acoustique) à sa réponse optimale [Benesty et al., 2007, chap 45]. Ce temps ne dépend pas seulement de la longueur du filtre adaptatif mais également du pas d'adaptation. Un pas d'adaptation élevé va permettre une adaptation très rapide. Un pas d'adaptation faible quant à lui mènera a une erreur estimation plus faible. Il y a un compromis à trouver au niveau de la valeur du pas d'adaptation. Il est très courant d'utiliser un pas d'adaptation variable [Benesty et al., 2006, Iqbal and Grant, 2008, Lee et al., 2009].

Le comportement de l'annulation d'écho adaptative dépend aussi de la présence de perturbations (signal utile, bruit ambiant, écho non-linéaire). Pour toutes ces raisons, le signal en sortie de l'annulation d'écho adaptative contient de l'écho résiduel. Il convient alors d'utiliser des modules supplémentaires afin de supprimer complètement l'écho :

- Le post-filtre qui a pour objectif de supprimer l'écho résiduel.

- Le modèle de l'annulation d'écho adaptative suppose que le couplage entre HP et le microphone est un phénomène linéaire. Ce qui n'est pas le cas, il existe des techniques spécifiques pour traiter l'écho non-linéaire [Birkett and Goubran, 1995b]. Elles consistent en consistent à utiliser un préprocesseur en amont du filtre adaptatif [Guerin et al., 2003, Stenger and Rabenstein, 1998] ou a rendre linéaire la réponse du HP [Furuhashi et al., 2006, Mossi et al., 2011].

A.3.1.2 Le post-filtre

L'approche la plus simple pour supprimer l'écho résiduel consiste en simplement atténuer le signal d'erreur provenant de l'annulation d'écho adaptative. Le gain d'atténuation est calculé en fonction du niveau du signal HP [Degry and Beaugeant, 2008, Heitkamper and Walker, 1993]. Mais cette solution mène à des mauvaises performances en périodes de double-parole. La voix du locuteur local sera alors partiellement atténuée - ce qui dégrade la qualité de la communication. Les techniques plus évoluées de suppression d'écho résiduel

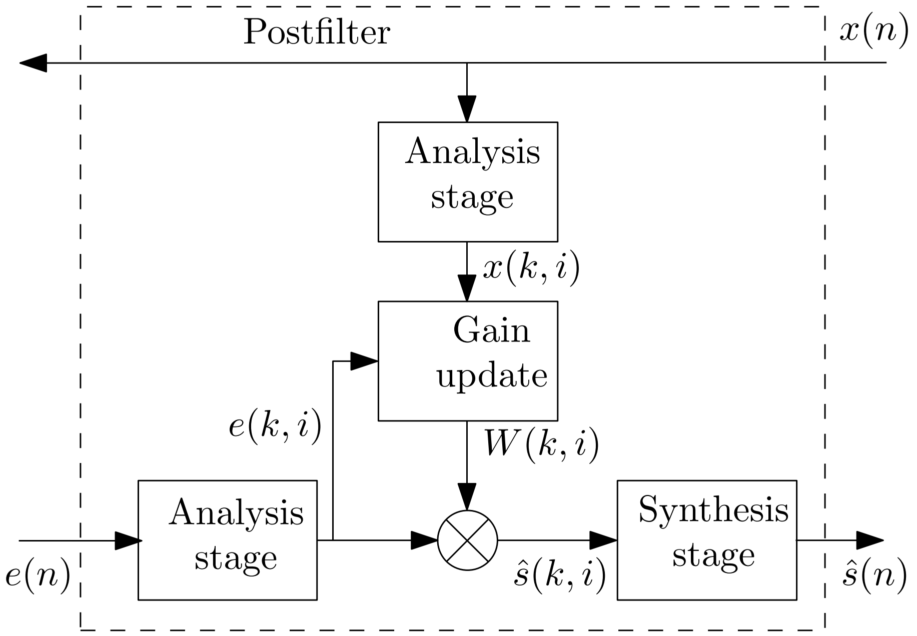


Figure 5: Schéma d'un post-filtre en sous-bande

consiste en utiliser un post-filtre dans le domaine fréquentiel ou sous-bande. Le principe d'un post-filtre en sous-bandes est illustré dans la Figure 5. Le principe est de calculer et d'appliquer le gain de suppression d'écho résiduel dans le domaine en sous-bande [Allen and Rabiner, 1977, Crochiere and Rabiner, 1983]. Le traitement en sous-bande permet de cibler les bandes de fréquences (ou fréquences) où il y a de l'écho résiduel. Les bandes de fréquences ne contenant que le signal utile ne seront pas atténuées : ce qui permet une meilleure conservation de la parole utilise en période de double-parole.

Il existe une variété de règles de calcul du gain en sous-bandes [Beaugeant et al., 1998, Ephraim and Malah, 1985]. Toutes ces règles vont avoir des paramètres qui vont permettre d'obtenir plus ou moins de d'atténuation d'écho. Les paramètres de réglages les plus courants sont :

- les facteurs de pondération d'estimation : ils permettent de prendre en compte les erreurs d'estimation. En fonction des besoins, il peut s'agit d'un facteur d'atténuation (pour réduire une estimée) ou d'un facteur d'amplification (pour augmenter une estimée). Le cas le plus courant est celui où ce facteur est utilisé pour sur-estimer l'estimée de la puissance de l'écho.

- constantes de lissage : elles vont permettre que le calcul gain prenne en compte une fenêtre d'observation plus ou moins longue. Ce facteur va influer la réactivité du post-filtre

A.3.1.3 L'approche synchronisée

L'annulation d'écho adaptative et le post-filtre ont pour but commun de supprimer l'écho : il conviendrait donc de pouvoir synchroniser les deux modules afin d'obtenir de meilleures performances. Il existe dans la littérature quelques techniques d'annulation d'écho dites synchronisées. L'objectif recherché via cette synchronisation est d'avoir une logique intelligente qui permette d'avoir un:

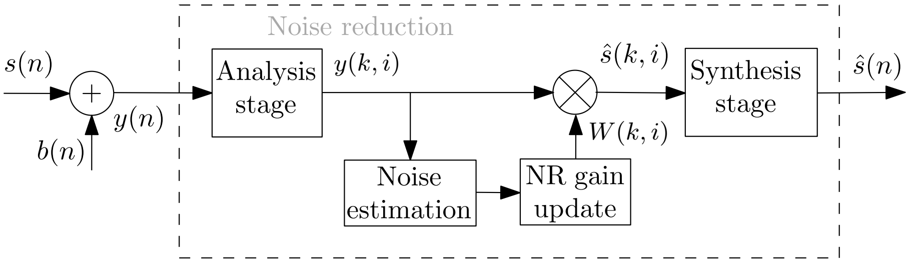


Figure 6: Schéma d'une ŕeduction de bruit

- Post-filtre plus agressif pendant les périodes de convergence de l'annulation d'écho adaptative.

- Post-filtre moins agressif lorsque l'annulation d'écho adaptative a atteint sa réponse optimale.

On distingue notamment deux techniques principales d'annulation d'écho synchronisée :

- Celles basées sur un modèle statique du canal acoustique : Ces approches exploitent le lien entre le pas d'adaptation de l'annulation d'écho adaptative et le post-filtre écho [Enzner and Vary, 2003, Steinert et al., 2007]. La contrainte de ces approches est que l'annulation d'écho adaptation et le post-filtre doivent être dans le même domaine en sous-bande ou fréquentiel.

- Celles basées sur un modèle dynamique du canal acoustique : L'idée est de modéliser les petites variations (dues au mouvement de la tête, mouvement d'un objet dans l'environnement acoustique) du canal acoustique par un modèle de Markov [Enzner and Vary, 2006, Haykin, 2002]

  $$h(n+1) = A \cdot h(n) + \Delta h(n) \tag{7}$$

  où A est le coefficient de transition [Enzner and Vary, 2006]. $\Delta h(n)$ modélise l'impredictibilité du canal acoustisque. En utilisant ce modèle le filtre adaptatif devient alors :

  $$\hat{H}(k+1,i) = A \cdot \hat{H}(k,i) + K(k,i) \cdot e(k,i) \tag{8}$$

  où $\hat{H}(k,i)$ est l'estimée de la transformée de Fourier de $h(n)$ et $K(k,i)$ est le gain de Kalman. La synchronisation vient de la méthode de calcul du gain de Kalman et du calcul du gain du post-filtre. Cette technique de synchronisation est plus robuste aux variations du canal acoustique. Mais son importante complexité de calcul est son principal inconvénient à son utilisation sur un téléphone portable.

A.3.2 La réduction de bruit

Nous avons expliqué que la qualité vocale peut aussi être dégradée par la présence de bruit ambiant. Dans la pratique les terminaux mobiles sont également équipés d'algorithmes de réduction de bruit. La Figure 6 montre le principe d'un module de réduction de bruit.

La réduction de bruit s'opère généralement dans le domaine en sous-bandes ou fréquentiel. Le signal bruité est utilisé pour obtenir une estimée du signal de bruit. Cette estimée

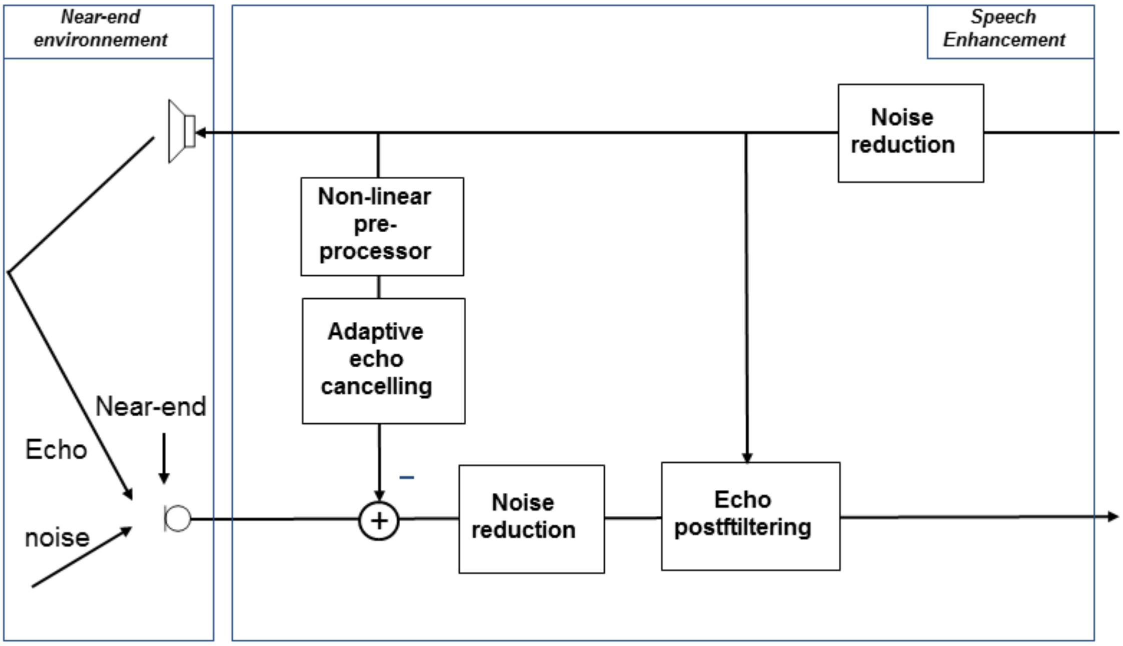


Figure 7: Exemple de chaîne de traitement de la parole pour terminal à un microphone.

est utilisée pour atténuer calculer le gain de réduction de bruit. Enfin, ce gain est appliqué au signal bruité afin d'atténuer le bruit. L'efficacité d'algorithme de réduction de bruit dépend principalement du choix du la méthode de calcule gain et de la méthode d'estimation du bruit. Parmi les méthodes de calcul de gain, on peut citer le gain de Wiener ou encore la soustraction spectrale. En ce qui concerne l'estimation du bruit, on distingue deux familles d'estimation :

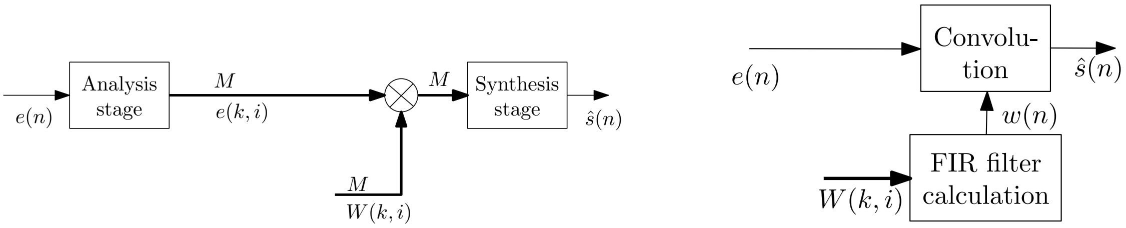
- Celles basées sur la détection des périodes d'activité vocales : ces techniques consistent en estimer le bruit uniquement en période de silence.

- Celles basées sur une estimation continue du niveau de bruit.

A.3.3 Notre objectif

Dans les sections A.3.1 et A.3.2, nous avons présenté quelques algorithmes de traitement de la parole de l'état de l'art. Dans la pratique, certains algorithmes peuvent être combinés afin d'obtenir une chaine de traitement de la parole qui permette d'améliorer la qualité vocale des communications téléphoniques. La Figure 7 montre un exemple de chaîne de traitement de la parole. Le préprocesseur placé avant le filtre adaptatif permet d'estimer les non-linéarités générées par le HP du terminal. De cette manière, l'annulation d'écho adaptative donnera une estimation du signal d'écho linéaire et non-linéaire. Le post-filtre quant à lui permet de supprimer l'écho résiduel qui subsiste à la sortie de l'annulation d'écho adaptative. Les modules de réduction de bruit permet d'atténuer les bruits tant pour le locuteur lointain que pour le locuteur local.

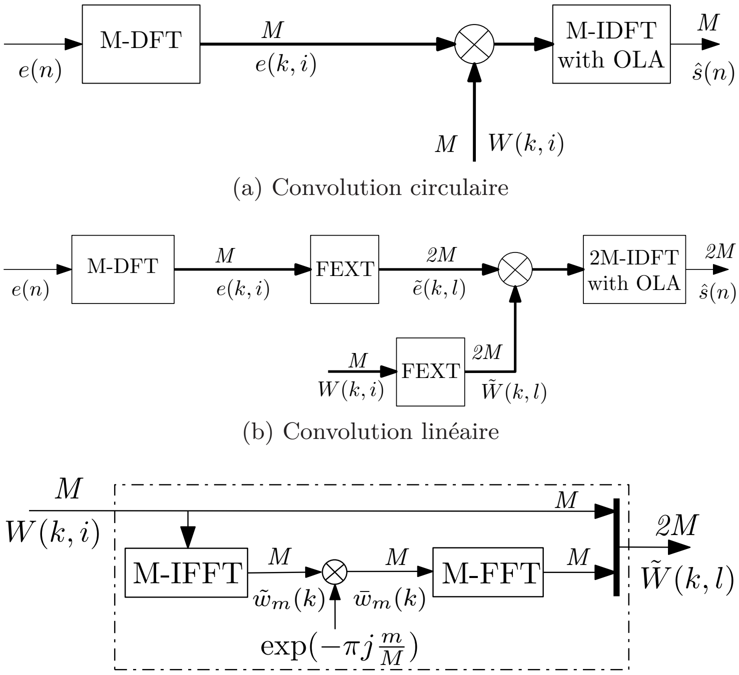
Cette thèse présente une nouvelle architecture combinée dâĂŹannulation d'écho pour terminaux mobiles à un ou deux microphones. LâĂŹarchitecture proposée réduit efficacement la complexité de calcul tout en améliorant la qualité de la parole dans les scénarios



(a) Filtrage dans le domaine en sous-bandes

(b) Filtrage dans le domaine temporel

Figure 8: Méthodes de filtrages. Les lignes en gras représentent le domaine en sous-bandes.



(a) Convolution circulaire

(b) Convolution linéaire

(c) Solution proposée: FEXT

Figure 9: Filtrage dans le domaine fréquentiel. Les lignes en gras représentent le domaine en fréquentiel.

défavorables. Nous présentons également la premiÃÍre solution bi-microphones de détection de double parole. Enfin, nos techniques bi-microphones peuvent facilement être appliquées aux terminaux multi-microphones et tout en ayant une complextité de calcul acceptable pour les téléphones mobiles.

A.4 Annulation d'écho pour terminaux à un microphone

A.4.1 La réduction conjointe de bruit et d'écho

Les techniques d'annulation d'écho de l'état de l'art sont basées sur une filtre adaptatif suivie par post-filtre. L'annulation d'écho adaptative vise à estimer le signal d'écho capté par le microphone tandis que le post-filtre consiste à atténuer l'écho résiduel dans le domaine fréquentiel ou en sous-bandes. De même, les approches de réduction de bruit pour but d'atténuer le bruit dans le domaine fréquentiel ou de sous-bande. Au vue des similarités entre le post-filtre et la reduction de bruit, une approche simple pour réduire la complexité de calcul de la chaîne de traitement de la parole consiste à combiner ces deux modules. Cette combinaison permet aussi de reduire le retard introduit par la chaîne de traitement de la parole

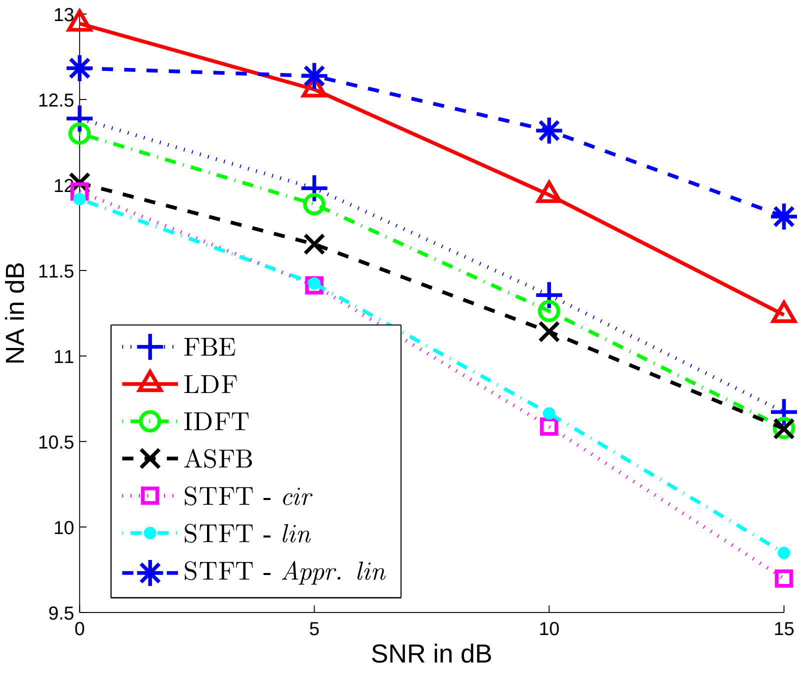
Le principe du systeme combiné se présente comme illustré sur la Figure 8 (a). En effet, en combinant les deux modules, nous sommes en mesure de réduire le retard introduit par les étapes d'analyse et syntheses puisque le système combine n'utilise qu'une analyse/synthèse au lieu de deux. Le retard introduit par le système combiné peut être davantage réduit en filtrant les perturbations (bruit ambiant et écho résiduel) dans le domaine temporel avec un filtre à réponse impulsionnelle fini (RIF). La Figure 8 (b) montre le schéma du système combine dans le cas où le filtrage a lieu dans le domaine temporel. Ce filtre RIF est calculé à partir des gains spectraux suivant la méthode du filter bank equalizer (FBE), du filtre à retard réduit LDF (LDF pour Low delay filter) ou de la transformée de Fourier discrète inverse (TFDI) [Löllmann and Vary, 2007, **Yemdji** et al., 2010a]. Le principal inconvénient des methodes de filtrage dans le domaine temporel est que le filtrage du signal se fasse par convolution, ce qui implique une augmentation de la complexité de calcul.

Nous comparons les performances de ces méthodes de filtrage temporel avec l'atténuation spectrale classique. En ce qui concerne l'atténuation spectrale, nous considérons deux cas : le cas où l'analyse et la synthèse sont effectuées par un banc de filtres [Hänsler and Schmidt, 2004] et le cas où l'analyse et la synthèse spectrales sont effectuées par transformée de Fourier à court terme (TFCT) avec chevauchement [Plapous, 2005]. Le cas avec la TFCT nous permet notamment d'évaluer la convolution linéaire et la convolution circulaire. La convolution circulaire souffre de repliement temporel: ce qui dans le cas de parole peut se traduire par des distorsions du signal utile. La convolution linéaire permet d'éviter ce type de distorsions mais requiert une complexité de calcul plus importante que la convolution circulaire. Nous proposons une solution alternative qui est basée sur une extension de la résolution fréquentielle hybride (Voir Figure 9 (c)). Cette solution a complexite de calcul plus faible que la convolution linéaire. Elle présente une scalabilité qui permet d'affiner encore plus la complexité de calcul.

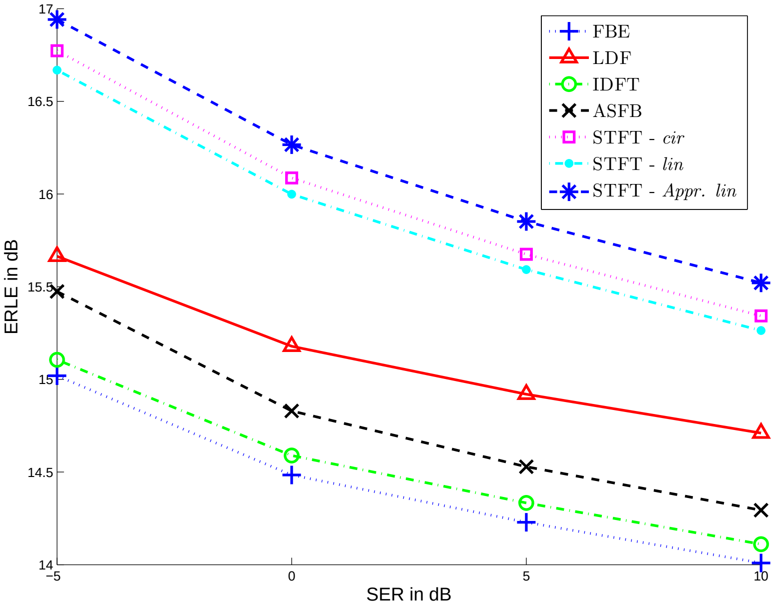
A.4.1.1 Observations

En somme notre étude porte sur la meilleure manière de combiner la réduction de bruit et le post-filtre. On compare notamment l'intéret de différentes méthode de filtrage: FBE, LDF, TFDI(ou encore IDFT), TFCT cir (ou encore STFT cir), TFCT lin (ou encore STFT lin) et le FEXT (ou encore STFT Appr.). Les mesures quantitatives de performances montrent que toutes les méthodes étudiées sont équivalentes en terme d'attenuation du bruit et de suppression d'écho (voir Figure 10).

Néanmoins, les écoutes informelles revèlent la présence de deux types de distorsions dans les signaux traités. Les signaux filtrés dans le domaine temporel sont altérés par des distorsions qui s'apparentent à un grésillement [**Yemdji** et al., 2010a]. On note tout de même que le niveau des artéfacts introduits par la méthode IDFT est légèrement supérieur à ceux des autres méthodes de filtrage temporel. Des distorsions de type bruit musical ont été observées pendant les périodes de double-parole dans les signaux issus de la TFCT *lin* : ceci est dû au fait que certaines composantes du signal de parole utile sont complètement atténuées en double-parole. Les signaux issus de la TFCT *cir* contiennent du bruit musical et parfois des grésillements, ces derniers sont observables pendant les périodes de double-



(a) Atténuation de bruit mésurée en période de parole utile seule



(b) Atténuation d'écho mésurée en période d'écho seul

Figure 10: Atténuation des pertubations.

parole. Le grésillement observé est la conséquence du repliement temporel inhérent à la convolution circulaire. Pour toutes les méthodes de filtrage étudiées, le niveau des artéfacts

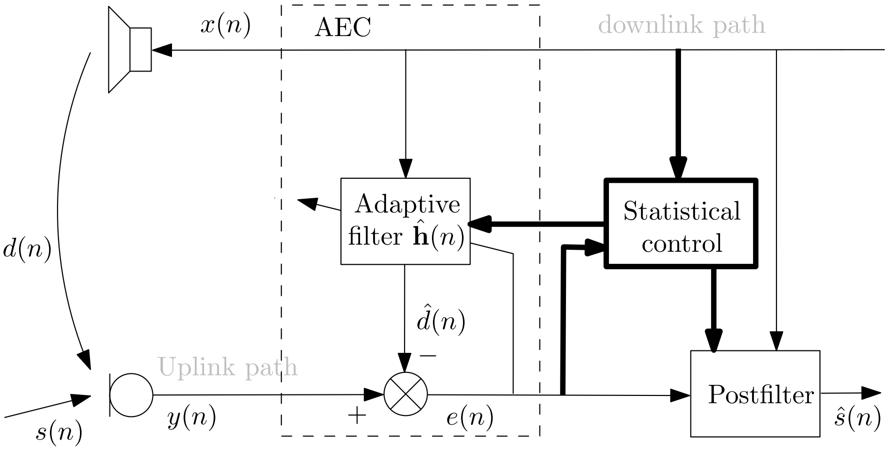


Figure 11: Principe de l'annulation d'écho par l'approche synchronisée

observés augmente quand les niveaux d'écho et/ou de bruit augmentent. Néanmoins, ces artéfacts ne sont pas perçus comme étant gênant dans la mésure où ils sont partiellement masqués par le bruit résiduel et/ou la parole utile. La parole reste audible dans tous les cas. Ces conclusions montrent que les techniques de fitrage temporel étudiées peuvent être efficacement utilisées à la place des l'atténuation spectrale classique.

A.4.2 L'annulation d'écho par l'approche synchronisée

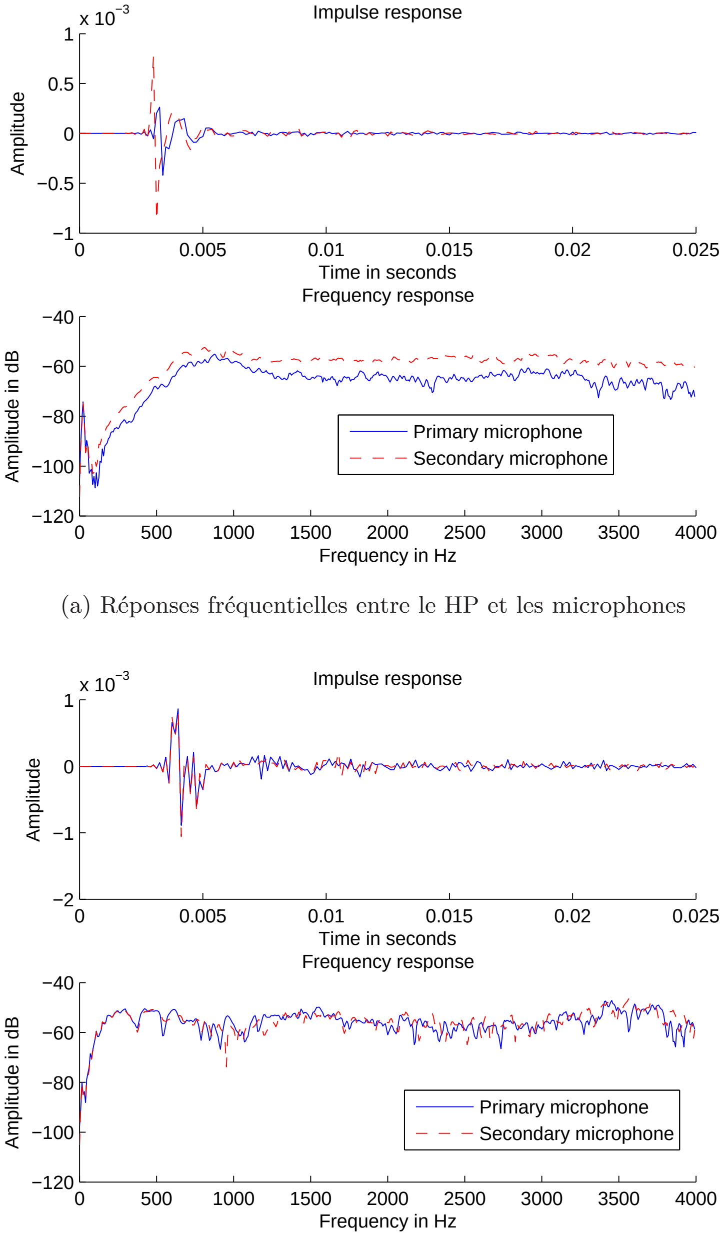
Un autre aspect abordé dans cette thèse est L'annulation d'écho par l'approche synchronisée. Notre approche pour synchroniser AEC et postfiltrage écho est basée sur un lien mathématique que nous établissons entre l'annulation d'echo adaptative et le post-filtre. La chaîne de traitement de l'écho se schématise alors comme illustrée sur la Figure 11. Le nouveau système necessite un module supplementaire qui permet de synchroniser l'annulation d'écho adaptative et le postfiltre. Les évaluations montrent les résultats de la méthode proposée de synchronisation dans la réduction significative du temps de convergence de l'AEC et de suppression meilleur écho .

A.5 Annulation d'écho pour terminaux à deux microphones

A.5.1 Problématique

Historiquement les techniques d'annulation d'écho utilisent un microphone. Une multitude d'algorithmes d'annulation d'écho adaptive et de post-filtre ont été developés autour de ces systèmes à un microphone. Néammoins, on constate que ces solutions à un microphone ont des performances limitées lorsque le niveau des pertubations (écho et bruit ambiant) augmente. Afin de pallier à ce type de limitations, on voit de plus en plus d'études de systèmes d'annulation d'écho avec des architectures multi-microphones. Il s'agit notamment des systèmes de formation de voies (ou beamforming). Les techniques existantes de beamforming utilisent des antennes qui ont entre 4 et 10 microphones, la distance entre les microphones quant à elle varie de 10cm à 1m.

Bien que les téléphones soient traditionnellement conȘ avec un microphone, on voit néanmoins apparaître quelques terminaux équipés de deux microphones. C'est le cas du *Google Nexus One*, de l'*iPhone 4* and des series S de chez *Samsung*. Quelques



(a) Réponses fréquentielles entre le HP et les microphones

(b) Réponses fréquentielles entre la bouche artificielle et les microphones

Figure 12: Réponses fréquentielles mésurées. Le téléphone est placé devant la bouche artificielle. Cette mésure est effectué en mode mains libres dans une cabine acoustique.

études récentes montrent ainsi comment ces architectures à deux microphones permettent d'améliorer la réduction de bruit [Dörbecker and Ernst, 1996, Jeub et al., 2012]. En ce qui concerne l'annulation d'écho avec deux microphones, très peu d'études existent. Une partie de notre travail durant cette thèse a portée sur l'annulation d'écho pour les terminaux mobiles à deux microphones. Ce travail s'est fait en deux temps:

- Dans un premier temps, nous avons effectué des mésures afin d'analyser les particularités de l'écho acoustique sur les terminaux à deux microphones. Nos mésures ont été effectue sur deux terminaux placés dans plusieurs environments acoustisques. Comme nous le montrons en section A.5.2.1, on observe qu'en periode d'écho seul, il existe une difference d'énergie importante entre les deux signaux microphoniques.

- Dans un second temps, nous proposons plusieurs solutions permettant d'améliorer l'annulation d'écho. Un synthèse des solutions proposées est présentée dans la section A.5.2.2.

A.5.2 Solutions proposées

Dans la suite, le terme microphone primaire désigne le microphone le plus éloigné du HP et le microphone secondaire désigne le microphone le plus proche du HP.

A.5.2.1 Le problème de l'écho dans des terminaux à deux microphones

Afin de pouvoir proposer des solutions qui exploitent au mieux les caractéristiques de l'écho dans les cas des téléphones portables à deux microphones, nous effectuons des mésures de l'écho et du parole utile en modes combiné et mains-libres. Les mésures sont effectuées à l'aide d'un mannequin qui joue le rôle du locuteur local. La Figure 12 montre un exemple de réponses impulsionnelles mésurées. On observe qu'en période d'écho seul, il y'a une importante différence d'énergie entre les deux signaux microphoniques. Cette difference a également été observé en mode combiné. En ce qui concerne les périodes de parole utile seule, on observe:

- En mode mains-libres, les signaux microphoniques ont approximativement la même énergie.
- En mode combiné, le signal capté par le microphone primaire a toujours plus d'énergie que le signal du microphone secondaire.

Nous exploitons ces différences d'énergie pour proposer de nouvelles techniques d'ánnulation d'écho.

A.5.2.2 Solutions

L'objectif de notre étude est de proposer une méthode de traitement de l'écho qui soit adaptée aux terminaux à deux microphones et qui ait une complexité de calcul acceptable pour un téléphone portable. Nous proposons de modifier la chaîne de traitement de l'ècho comme illustré en Figure 13:

- L'annulation d'écho adaptative: tout comme dans le cas des terminaux à un microphone, elle va permettre d'estimer le signal d'écho. L'idée ici est de placer l'annulation d'écho adaptative sur le microphone primaire puisqu'il contient le moins d'écho.
- Le post-filtre: Contrairement aux techniques de la littérature, nous suggÃĺrons d'utiliser les deux signaux microphones pour la supprśsion d'écho résiduel. La suppression d'écho résiduel continue de s'effectuer par atténuation spectrale du signal d'erreur en sortie de l'annulation d'écho adaptative. Le principal intéret de notre technique réside en le fait que le calcul du gain en sous bande utilise deux microphones (au lieu d'un). Il convient donc de concevoir ou proposer de nouvelles règles de calcul du gain.

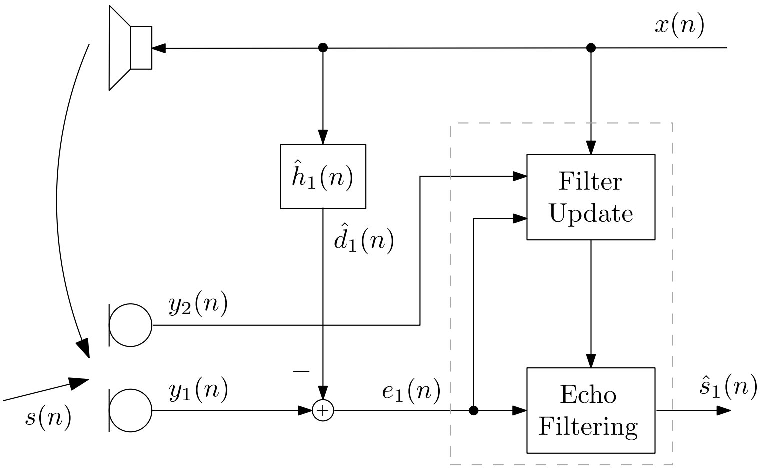


Figure 13: Annulation d'écho pour terminal à deux microphones

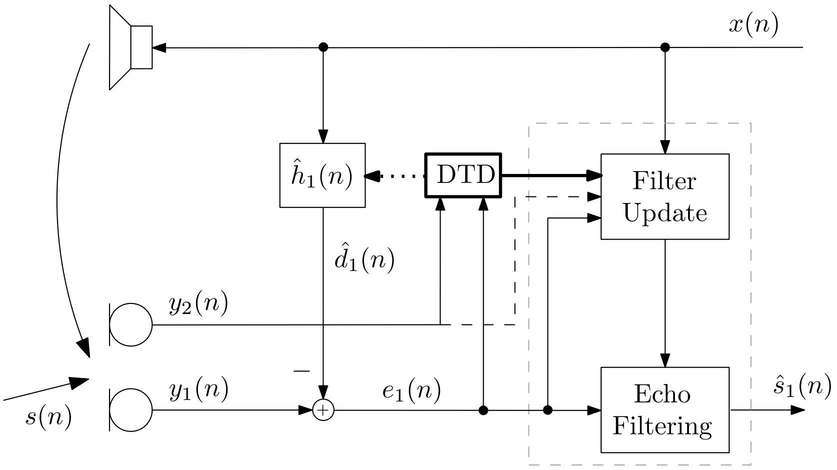


Figure 14: DM echo cancellation scheme including PLD based DTD control

Dans cette thèse, nous introduisons deux nouvelles règles de calcul de gain toutes deux derivées de la règle de Wiener:

- La premiere méthode exploite la diffeérence d'énergie observée entre les signaux microphoniques en période d'écho seul. Une étude mathématique de cette nouvelle règle de calcul montre qu'elle a des performances limitées en périodes d'écho seul.

- La deuxième exploite la corrélation entre les signaux microphones. Elle présente l'intérêt de pouvoir être étendue de manière càibler aussi l'écho non-linéaire.

Par ailleurs, nous exploitons la différence d'énergie observée en echo seul pour définir un détecteur de double-parole. Ce détecteur est basée sur une mésure de la différence d'énergie entre les deux microphones. La détection peut tout aussi bien s'effectuer dans le domaine frqentiel que le domaine temporel. Dans notre cas, nous avons effectué la detection dans le domaine fréquentiel. De cette manière, ce détecteur peut être utilisé pour contrôler l'annulation d'écho adaptative et le post-filtre. Dans ce cas la chaîne de traitement de l'écho se présente comme illustrée en Figure 14.

A.5.3 Observations et conclusions

La problématique de cette partie de la thèse concerne le contrôle de l'écho pour les téléphones portables à deux microphones. Une analyse du problème d'écho basée sur des enregistrements est proposée. Il résulte de cette étude qu'en placant les transducteurs dans une certaine configuration, on observe d'importantes différences d'énergies entre les signaux microphoniques selon que nous sommes en écho seul, parole utile seule ou double parole.

Contrairement aux systèmes existants de contrôle d'écho multi- microphone, nous choisissons d'utiliser un filtre adaptatif suivie par post-filtrage. Dans le chapitre 5, nous montrons comment la différence de niveau observée avec nos enregistrements peut être exploitée pour la détection de double-parole. Deux nouveaux post-filtres sont également proposés. Notre évaluation montre que le détecteur de double-parole proposé peut être utilisé efficacement pour améliorer n'importe quel algorithme existant de contrôle d'écho à un microphone. Les experiences montrent l'utilisation de notre détecteur de double-parole permet d'améliorer les performances du post-filtre à un microphone: on note une augmentation de la quantité d'écho supprimé et une baisse du niveau de distortion en période de double-parole. Les écoutes informelles montrent que cette baisse du niveau de distorsion est perceptible. Dans le monde industriel, il est très difficile d'effectuer un changement radical d'une solution existante. Notre détecteur de double parole peut être vue comme une étape vers la mise en place d'un post-filtre utilisant deux microphones.

Dans nos experiences, le post-filtre basé sur la différence d'énergie est utilisé en combinaison avec notre détecteur de double-parole. Les expériences montrent que le post-filtre ainsi obtenu permet une amélioration notable des performances en terme de l'annulation d'écho et de distorsions introduites. En periode de double-parole les distorsions sont quasiment inaudibles. On observe également que les performances de notre annulation d'écho augmentent lorsque la différence d'énergie entre les signaux microphones augmentent. Ce qui nous permet d'émettre une recommendation concernant le positionnement des transducteurs sur les téléphones portables. **Le microphone primaire devrait être placée aussi loin que possible du HP tandis que le microphone secondaire doit Ãłtre placé aussi près que possible du HP afin de garantir une différence d'énergie maximale entre les deux signaux.**

Bibliography

J. Allen and L. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558 – 1564, nov. 1977.

L. Azpicueta-Ruiz, M. Zeller, A. Figueiras-Vidal, J. Arenas-Garcia, and W. Kellermann. Adaptive combination of Volterra kernels and its application to nonlinear acoustic echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):97 –110, Jan. 2011.

F. Balle. *Médias et sociétés: presse, édition, Internet, radio, cinéma, télévision, télématique, cédéroms, DVD, réseaux multimédias*. Domat Politique. Montchrestien, 2005.

C. Beaugeant. *Réduction de bruit et contrôle de l'écho pour les applications radiomobiles*. PhD thesis, Université de Rennes 1, 1996.

C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire. New optimal filtering approaches for hands-free telecommunication terminals. *Signal Processing*, 64(1):33–47, 1998.

A. Benamar. *Etude et implantation de la fonction de contrôle de l'écho acoustique pour la radiotéléphonie mains-libres*. PhD thesis, Université de Paris-Sud, Orsay, 1996.

J. Benesty, H. Rey, L. Vega, and S. Tressens. A nonparametric VSS NLMS algorithm. *IEEE Signal Processing Letters*, 13(10):581 –584, oct. 2006.

J. Benesty, M. M. Sondhi, and Y. A. Huang. *Springer handbook of speech processing*. Springer, 2007.

A. Birkett and R. Goubran. Acoustic echo cancellation using nlms-neural network structures. In *International Conference on Acoustics, Speech, and Signal Processing ICASSP*, volume 5, pages 3035–3038. IEEE, 1995a.

A. N. Birkett and R. A. Goubran. Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects. In *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio andAcoustics (WASPAA)*, pages 103 – 106, 1995b.

S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2):113–120, 1979.

H. Buchner, J. Benesty, T. Gansler, and W. Kellermann. Robust extended multidelay filter and double-talk detector for acoustic echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1633–1644, 2006.

O. Cappe. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *Speech and Audio Processing, IEEE Transactions on*, 2(2):345–349, 1994.

T. Claasen and W. Mecklenbrauker. Comparison of the convergence of two algorithms for adaptive FIR digitalfilters. *IEEE Trans. on Circuits and Systems*, 28(6):510 – 518, Jun 1981.

I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *Speech and Audio Processing, IEEE Transactions on*, 11(5):466–475, 2003.

R. E. Crochiere and L. R. Rabiner. *Multirate digital signal processing.* Prentice-Hall, 1983.

N. Curien and M. Gensollen. *Economie des télécommunications: Ouverture et réglementation.* Management - Communication - Réseaux. Economica, 1992.

A. Davis, S. Nordholm, and R. Togneri. Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):412–424, 2006.

P. Degry and C. Beaugeant. Solution to speech quality improvement in telecommunication terminals. In *Proc. ITG Fachtagung Sprachkommunikation*, Oct. 2008.

M. Dörbecker and S. Ernst. Combination of two-channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation. In *Proc. European Signal Processing Conference (EUSIPCO*, 1996.

D. Duttweiler. A twelve-channel digital echo canceler. *Communications, IEEE Transactions on*, 26(5):647–653, 1978.

G. Enzner. *A model-based optimum filtering approach to acousti echo control: Theory and proactice.* PhD thesis, Institut für Nachrichtengeräte und Datenverarbeitung (IND), RWTH Aachen, 2006.

G. Enzner and P. Vary. Robust and elegant, purely statistical adaptation of acoustic echo cancelerand postfilter. In *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, Sep. 2003.

G. Enzner and P. Vary. Frequency-domain adaptive kalman filter for acoustic echo control in hands-freetelephones. *Signal Processing*, 86(6):1140 – 1156, 2006.

G. Enzner, R. Martin, P. Vary, G. Enzner, R. Martin, and P. Vary. Partitioned residual echo power estimation for frequency-domain acoustic echo cancellation and postfiltering. *European transactions on telecommunications*, 13(2):103–114, 2002.

Y. Ephraim and D. Malah. Speech enhancement using optimal non-linear spectral amplitude estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, pages 1118–1121, Apr. 1983.

Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):443–445, Apr. 1985.

A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*, 2 2000. URL `http://www.aes.org/e-lib/browse.cfm?elib=10211`.

T. Fingscheidt and S. Suhahi. Quality assessment of speech enhancement systems by separation of enhanced speech, noise and echo. In *Proc. Interspeech*, pages 818 – 821, Antwerp, Belgium, 2007.

P. Flichy. *Dynamics of modern communication: the shaping and impact of new communication technologies.* Media, culture, and society series. Sage Publications, 2004.

H. Furuhashi, Y. Kajikawa, and Y. Nomura. Realization of nonlinear acoustic echo cancellation by subband parallel cascade volterra filter. In *Intelligent Signal Processing and Communications, 2006. ISPACS '06. International Symposium on*, pages 837–840, 2006.

S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614 –1626, aug 2001.

T. Gänsler and J. Benesty. A frequency-domain double-talk detector based on a normalized cross-correlation vector. *Signal Processing*, 81(8):1783 – 1787, 2001. Special section on Signal Processing Techniques for Emerging Communications Applications.

T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson. A double-talk detector based on coherence. *Communications, IEEE Transactions on*, 44(11):1421–1427, 1996.

F. X. Gao and W. M. Snelgrove. Adaptive linearization of a loudspeaker. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3589–3592. IEEE, 1991.

T. Gerkmann and R. Hendriks. Noise power estimation based on the probability of speech presence. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 145–148, 2011.

M. Goulding and J. Bird. Speech enhancement for mobile telephony. *IEEE Transactions on Vehicular Technology*, 39(4):316 –326, nov 1990.

A. Guerin. *Rehaussement de la parole pour les communications mains-libres. Réduction de bruit et annulation dŠécho non linéaire parlée.* PhD thesis, Thèse de l'Université de Rennes 1, 2002.

A. Guerin, G. Faucon, and R. Le Bouquin-jeannes. Nonlinear acoustic echo cancellation based on volterra filters. *Speech and Audio Processing, IEEE Transactions on*, 11(6): 672–683, 2003.

M. Guo, T. Elmedyb, S. Jensen, and J. Jensen. Acoustic feedback and echo cancellation strategies for multiple-microphone and single-loudspeaker systems. In *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 556 –560, nov. 2011.

H. Gustafsson, S. Nordholm, and I. Claesson. Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Trans. on Speech and Audio Processing*, 9 (8):799 –807, Nov. 2001.

S. Gustafsson, R. Martin, P. Jax, and P. Vary. A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. *Speech and Audio Processing, IEEE Transactions on*, 10(5):245–256, 2002.

E. Habets, S. Gannot, I. Cohen, and P. Sommen. Joint dereverberation and residual echo suppression of speech signals innoisy environments. *IEEE Trans. on Audio, Speech, and Language Processing*, 16(8):1433 – 1451, Nov. 2008.

E. A. P. Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technische Universiteit Eindhoven, 2007. URL `http://alexandria.tue.nl/extra2/200710970.pdf`.

E. Hänsler and G. Schmidt. Hands-free telephones - joint control of echo cancellation and postfiltering. *Signal Processing*, 80(11):2295–2305, 2000.

E. Hänsler and G. Schmidt. *Acoustic Echo and Noise Control: A Practical Approach*. Wiley-Interscience, 2004.

S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 2002.

HEAD Acoustics HMS II.3. Head measurement system with ear simulator and mouth simulator. URL `http://www.head-acoustics.de/eng/telecom_hms_II_3.htm`.

P. Heitkamper and M. Walker. Adaptive gain control for speech quality improvement and echo suppression. In *Proc. IEEE International Symposium on Circuits and Systems*, pages 455–458, May 1993.

O. Hoshuyama. An update algorithm for frequency-domain correlation model in a nonlinear echo suppressor. In *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC) 2012*, pages 1 –4, sept. 2012.

Y. Huang, J. Benesty, and J. Chen. *Acoustic MIMO signal processing*. Springer Berlin, Germany, 2006.

J. Huo, S. Nordholm, and Z. Zang. New weight transform schemes for delayless subband adaptive filtering. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pages 197 –201 vol.1, 2001.

M. Iqbal and S. Grant. Novel variable step size NLMS algorithms for echo cancellation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 241–244, 2008.

ITU-T. ITU-T recommendation P.56: objective measurement of active speech level, 1993.

ITU-T. ITU-T recommendation P.340: Transmission characteristics of handsfree telephones, 1996a.

ITU-T. ITU-T recommendation P.800: Methods for objective and subjective assessment of quality, 1996b.

ITU-T. ITU-T recommendation P.64: Determination of sensitivity/frequency characteristics of local telephone systems, 2007.

W. Jeannes, P. Scalart, G. Faucon, and C. Beaugeant. Combined noise and echo reduction in hands-free systems: a survey. *IEEE Transactions on Speech and Audio Processing*, 9 (8):808 –820, nov 2001.

M. Jeub, C. Nelke, H. Krüger, C. Beaugeant, and P. Vary. Robust dual-channel noise power spectral density estimation. In *Proc. European Signal Processing Conference (EUSIPCO)*, pages 2304–2308, Barcelona, Spain, Aug. 2011.

M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary. Noise reduction for dual microphone mobile phones exploiting power level differences. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, pages 1693–1696, Kyoto, Japan, Mar. 2012.

W. Kellermann. Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 219 –222, apr 1997.

M. Krawczyk and T. Gerkmann. STFT phase improvement for single channel speech enhancement. In *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, pages 1–4, 2012.

H. Kuttruff. *Room Acoustics.* E-Libro. Taylor & Francis, 2000.

M. Leclerc and P. Carré. *France Télécom: mémoires pour l'action.* France Télécom, 1995.

J. Lee, J.-W. Chen, and H.-C. Huang. Performance comparison of variable step-size nlms algorithms. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1, pages 20–22, 2009.

J. Lim and A. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

H. W. Löllmann. *Allpass-Based Analysis-Synthesis Filter-Banks: Design and Application.* PhD thesis, Universitätsbibliothek, 2011.

H. W. Löllmann and P. Vary. Uniform and warped low delay filter-banks for speech enhancement. *Speech Communications*, 49(7–8):574–587, 2007.

H. W. Löllmann and P. Vary. A blind speech enhancement algorithm for the suppression of late reverberationand noise. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, pages 3989–3992, 2009.

S. Malik and G. Enzner. Model-based vs. traditional frequency-domain adaptive filtering in the presenceof continuous double-talk and acoustic echo path variability. In *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

J. Marin-Hurtado and D. Anderson. Distortions in speech enhancement due to block processing. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, pages 4774 –4777, 2010.

R. Martin. Combined acoustic echo cancellation, spectral echo shaping, and noise reduction. In *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 48–51, 1995.

R. Martin. Noise power spectral density estimation based on optimal smoothing and minimumstatistics. *IEEE Trans. on Speech and Audio Processing*, 9(5):504 – 512, Jul. 2001.

R. Martin and J. Altenhoner. Coupled adaptive filters for acoustic echo control and noise reduction. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 3043 –3046 vol.5, may 1995.

D. Morgan and J. Thi. A delayless subband adaptive filter architecture. *IEEE Transactions on Signal Processing*, 43(8):1819 –1830, aug 1995.

M. Mossi, N. W. D. Evans, and C. Beaugeant. An assessment of linear adaptive filter performance with nonlinear distortions. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, 2010.

M. Mossi, C. Yemdji, N. W. D. Evans, and C. Beaugeant. Non-linear acoustic echo cancellation using online loudspeaker linearization. In *WASPAA 2011, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, UNITED STATES, 10 2011.

V. Myllyla and M. Hamalainen. Adaptive beamforming methods for dynamically steered microphone array systems. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 305–308, 2008.

A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, 1999.

C. Paleologu and J. Benesty. A practical data-reuse adaptive algorithm for acoustic echo cancellation. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2010–2014, 2012.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, Nov 2012. URL `http://www2.imm.dtu.dk/pubdb/p.php?3274`. Version 20121115.

C. Plapous. *Traitements pour la réduction de bruit. Application à la communication parlée.* PhD thesis, Thèse de l'Université de Rennes 1, 2005.

J. Proakis and D. Manolakis. *Digital signal processing: principles, algorithms, and applications.* Prentice-Hall International editions. Prentice Hall, 1996.

G. Reuven, S. Gannot, and I. Cohen. Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller. *Speech communication*, 49(7): 623–635, 2007a.

G. Reuven, S. Gannot, and I. Cohen. Multichannel acoustic echo cancellation and noise reduction in reverberant environments using the transfer-function gsc. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, 2007b.

P. Scalart and J. Filho. Speech enhancement based on a priori signal to noise estimation. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 629–632 vol. 2, 1996.

O. Shalvi and E. Weinstein. System identification using nonstationary signals. *IEEE Transactions on Signal Processing*, 44(8):2055–2063, 1996.

K. Shi, X. Ma, and G. Zhou. A residual echo suppression technique for systems with nonlinear acoustic echo paths. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 257–260, 2008.

B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan. A parametric formulation of the generalized spectral subtraction method. *Speech and Audio Processing, IEEE Transactions on*, 6(4):328–337, 1998.

P. Sommen and J. Jayasinghe. On frequency domain adaptive filters using the overlap-add method. In *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 27 –30, Jun. 1988.

K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt. Low-delay subband echo control in automotive environment. In *Proc. Biennial on DSP for in-Vehicule and Mobile Systems*, Istanbul, Turkey, Jun. 2007.

K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt. Hands-free system with low-delay subband acoustic echo control and noisereduction. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, pages 1521–1524, Mar. 2008.

A. Stenger and W. Kellermann. Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling. *Signal Processing*, 80(9):1747 – 1760, 2000.

A. Stenger and R. Rabenstein. An acoustic echo canceller with compensation of nonlinearities. In *Proc. EUSIPCO*, volume 98, pages 969–972, 1998.

S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, 8(4):478–482, 2000.

C. Yemdji, M. Mossi I., N. W. D. Evans, and C. Beaugeant. Efficient low delay filtering for residual echo suppression. In *Proc. European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, Aug. 2010a.

C. Yemdji, M. Mossi I., N. W. D. Evans, and C. Beaugeant. Low delay filtering for joint noise reduction and residual echo suppression. In *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel'Aviv, Israël, Sep. 2010b.

C. Yemdji, M. Mossi I., N. W. D. Evans, and C. Beaugeant. A scalable architecture for linear convolution in the frequency domain forspeech enhancement. In *Proc. Digital Signal Processing (DSP)*, 2011.

C. Yemdji, N. W. D. Evans, C. Beaugeant, and L. Lepauloux. Methods for processing audio signals and circuit arrangements therefor, Nov. 2012a. US 13/676142.

C. Yemdji, M. Mossi Idrissa, N. W. D. Evans, and C. Beaugeant. A new cross-domain approach to synchronized adaptive echo cancellation and echo postfiltering. In *Proc. European Signal Processing Conference, (EUSIPCO)*, Bucharest, Romania, Aug. 2012b.

C. Yemdji, M. Mossi Idrissa, N. W. D. Evans, C. Beaugeant, and P. Vary. Dual channel echo postfiltering for hands-free mobile terminals. In *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012c.

C. Yemdji, L. Lepauloux, C. Beaugeant, and N. W. D. Evans. Method for processing an audio signal and audio receiving circuit, May 2013. US 13/892,420.

V. Turbin, A. Gilloire, and P. Scalart. Comparison of three post-filtering algorithms for residual acoustic echo reduction. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 1, pages 307–310 vol.1, 1997.

P. P. Vaidyanathan. *Multirate systems and filter banks*. Pearson Education India, 1993.

P. Vary. Noise suppression by spectral magnitude estimation mechanism and theoretical limits. *Signal Processing*, 8(4):387 – 400, 1985.

D. Wang and J. Lim. The unimportance of phase in speech enhancement. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 30(4):679–681, 1982.

Xolo X900. URL `http://www.xolo.in/x900`.