LOW-COMPLEXITY ACOUSTIC ECHO CANCELLATION AND MODEL-BASED RESIDUAL ECHO SUPPRESSION

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von
Naveen Kumar Desiraju
geboren am 23. Oktober 1989
in Nagpur (Indien)

Naveen Kumar Desiraju: *Low-complexity acoustic echo cancellation and model-based residual echo suppression*

BETREUER:
Prof. Dr. ir. Simon Doclo, *University of Oldenburg, Germany*
Dr.-Ing. Tobias Wolff, *Cerence Inc., Ulm, Germany*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo, *University of Oldenburg, Germany*

WEITERE GUTACHTER:
Prof. Dr. ir. Emanuël Habets, *University of Erlangen-Nuremberg, Germany*
Prof. Dr.-Ing. Gerald Enzner, *University of Oldenburg, Germany*

TAG DER DISPUTATION:
18. Februar 2022

ACKNOWLEDGMENTS

This thesis has been written during my time as an external Ph.D. student at the Signal Processing Group in the Department of Medical Physics and Acoustics of the Carl von Ossietzky Universität Oldenburg in Oldenburg, Germany. I would like to take this opportunity to thank the many people who contributed in different ways to the completion of this work.

I would like to begin by expressing my gratitude to my supervisor Prof. Simon Doclo for providing support, advice and guidance, for his confidence in me and my work, and for the many interesting scientific and non-scientific discussions over the years. I thank him in particular for his detailed remarks and suggestions which always resulted in improving the quality of my work. A massive thank you to my co-supervisor Tobias Wolff for his support, guidance and friendship and for his unshakeable belief in me. Our constant discussions and exchange of ideas contributed immensely in shaping the course of my research. I am especially grateful to him for his patience and belief in me during a particularly difficult period of my academic journey, and for his constant encouragement which filled me with the confidence I needed to complete this work. I would like to additionally thank Markus Buck and Timo Gerkmann for the interesting discussions and insightful feedback. Furthermore, I would like to thank Prof. Emanuël Habets for taking the time to review my thesis and for his interest in my work, as well as Prof. Gerald Enzner as a further member of the examination committee.

I would like to thank my former colleagues and co-Ph.D. students Ingo Schalk-Schupp and Simon Graf, who were with me on this journey together from the start, and with whom I shared many memorable moments, both professional and personal. I would also like to thank my former teammates in the Signal Processing Group who shared this journey with me, especially Ante Jukić, Ina Kodrasi, Daniel Marquardt and Nico Gößling. A special thanks to the European Union's DREAMS project, which funded a majority of my research, and gave me the opportunity to travel to different universities across Europe and continuously meet with talented researchers.

Finally, I would like to thank my friends, in particular Vance and Harish, my brother Arun and my beloved wife Ishi, for their years of continuous support and encouragement. Last but not least, I would like to thank my father Dr. D.L.N. Rao and my mother Dr. P. Surekha Rao for their constant love and support, and for pushing me at every stage to reach my potential and not letting me rest in my comfort zone.

Nürnberg, October 2022

Naveen Kumar Desiraju

ABSTRACT

Hands-free speech communication devices, typically equipped with multiple microphones and loudspeakers, are used for a wide variety of applications, such as teleconferencing, in-car communication and personal assistants. In addition to capturing the desired speech from the user, the microphones pick up undesired interferences such as background noise and acoustic echo due to the acoustic coupling between the loudspeakers and the microphones. These interferences typically degrade speech quality and intelligibility, and negatively affect the performance of automatic speech recognition systems.

Acoustic echo control systems typically employ a combination of acoustic echo cancellation (AEC) and residual echo suppression (RES). An AEC system uses adaptive filters to compensate for the acoustic echo paths between the loudspeakers and the microphones. When short AEC filters are used to reduce computational complexity and increase convergence speed, this may lead to a significant amount of residual echo, which is typically suppressed using a RES postfilter. To compute the spectral weights of this postfilter, an accurate estimate of the power spectral density (PSD) of the residual echo is required.

The main aim of this thesis is to achieve low-complexity acoustic echo cancellation for multichannel systems by developing efficient tap selection schemes for partially updating the AEC filters, and to develop model-based residual echo PSD estimators for improved residual echo suppression.

First, we propose novel tap selection schemes which exploit input signal sparsity across the dimensions of frequency, channels and time, leading to efficient partial updates of multichannel AEC filters in the subband domain. In particular, the proposed dynamic effort allocation scheme proportionately selects more filter taps for update in subbands and channels with larger magnitude tap-inputs while not ignoring the filters with smaller magnitude tap-inputs. Simulation results for both synthetic as well as real-world multichannel input signals show that the proposed tap selection scheme achieves similar echo cancellation performance compared to updating all filter taps at a significantly reduced computational cost (about 28%).

Second, we propose novel signal-based methods to estimate the late residual echo PSD in online mode. The late residual echo PSD is modeled using an infinite impulse response (IIR) filter on the PSD of the loudspeaker signal, based on frequency-dependent reverberation scaling and decay parameters. We propose several signal-based methods based on output error and equation error to jointly estimate both reverberation parameters by minimizing a single cost function in online mode. Simulation results using both artificially generated as well as measured impulse responses

show that the output error method minimizing the mean squared log error (MSLE) cost function outperforms state-of-the-art offline and online methods in terms of parameter estimation accuracy, late residual echo PSD estimation accuracy and residual echo suppression performance.

Third, we propose a novel model for the early residual echo PSD and combine it with the IIR filter model for the late residual echo PSD to yield a novel model for the residual echo PSD. In particular, we model the early residual echo PSD using a moving average filter on the PSD of the loudspeaker signal, based on a frequency-dependent coupling factor. We propose signal-based methods based on output error to jointly estimate all three model parameters, i.e., the coupling factor and the reverberation scaling and decay parameters, by minimizing a single MSLE cost function in online mode. Simulation results using both artificially generated as well as measured impulse responses show that the proposed output error method with the recursive prediction error algorithm outperforms state-of-the-art offline and online parameter estimation methods in terms of parameter estimation accuracy and residual echo PSD estimation accuracy. Compared to state-of-the-art RES methods, the proposed method yields the best segmental speech-to-speech distortion ratio score (about 2-5 dB better), while also yielding the best segmental residual echo attenuation score (about 1-2 dB better).

ZUSAMMENFASSUNG

Freisprecheinrichtungen, die in der Regel mit mehreren Mikrofonen und Lautsprechern ausgestattet sind, werden für eine Vielzahl von Anwendungen eingesetzt, z. B. für Telefonkonferenzen, Kommunikation im Auto und persönliche Assistenten. Die Mikrofone nehmen dabei nicht nur die gewünschte Sprache des Benutzers auf, sondern auch unerwünschte Störungen wie Hintergrundgeräusche und Echos welche aufgrund der akustischen Kopplung zwischen den Lautsprechern und den Mikrofonen entstehen. Diese Störungen beeinträchtigen in der Regel sowohl die Sprachqualität als auch die Sprachverständlichkeit und wirken sich negativ auf die Leistung von automatischen Spracherkennungssystemen aus.

Systeme zur Reduktion akustischer Echos verwenden in der Regel eine Kombination aus akustischer Echokompensation (AEC) und Restechounterdrückung (RES). Ein AEC-System verwendet adaptive Filter, um die akustischen Echowege zwischen den Lautsprechern und Mikrofonen zu kompensieren. Wenn kurze AEC-Filter verwendet werden, um die Rechenkomplexität zu reduzieren und die Konvergenzgeschwindigkeit zu erhöhen, kann dies zu einer erheblichen Menge an Restechos führen, die normalerweise mit einem RES-Nachfilter unterdrückt werden. Um die spektralen Gewichte dieses Nachfilters zu berechnen, ist daher eine genaue Schätzung der spektralen Leistungsdichte (engl. „power spectral density“, PSD) des Restechos erforderlich.

Das Hauptziel dieser Arbeit ist es, eine akustische Echokompensation mit geringer Komplexität für Mehrkanalsysteme zu erreichen, indem neuartige Verfahren zur teilweisen Aktualisierung der AEC-Filterkoeffizienten sowie modellbasierte PSD-Schätzer für eine verbesserte Restechounterdrückung entwickelt werden.

Zunächst werden neuartige Auswahlverfahren für die AEC-Filterkoeffizienten vorgeschlagen, welche ausnutzen, dass die wiedergegebenen Lautsprechersignale über die Dimensionen Frequenz, Kanäle und Zeit oft dünn besetzt sind. Dies führt zu effizienten Teilaktualisierungen von Mehrkanal-AEC-Filtern im Teilbandbereich („sparse updates“). Das vorgeschlagene Verfahren zur dynamischen Verteilung des für das Filter-Update verfügbaren Aufwands verwendet die Resourcen zwar hauptsächlich in Bereichen großer Signalmagnituden, vermeidet es jedoch Bereiche mit geringen Signalmagnituden gänzlich zu vernachlässigen. Simulationsergebnisse für sowohl synthetische als auch reale Mehrkanal-Eingangssignale zeigen, dass dieses Verfahren bei deutlich reduzierten Kosten (ca. 28%) eine vergleichbare Güte der Echokompensation wie die vollständige Aktualisierung der Filterkoeffizienten erzielt.

Zweitens werden neuartige signalbasierte Methoden zur PSD-Schätzung des späten Restechos im Online-Modus vorgestellt. Das PSD des späten Restechos wird dabei mit einem Filter mit unendlich langer Impulsantwort (IIR) bestimmt welches auf

das PSD des Lautsprechersignals angewendet wird. Die Filterparameter werden hierbei frequenzabhängig optimiert. Es werden mehrere signalbasierte Verfahren zur Echtzeitschätzung der Parameter betrachtet, welche auf den Prinzipien des Ausgangs- und des Gleichungsfehlers beruhen und die Parameter durch Minimierung einer gemeinsamen Kostenfunktion bestimmen. Simulationsergebnisse, die sowohl auf künstlich erzeugten als auch auf gemessenen Impulsantworten basieren, zeigen, dass die Ausgangsfehlermethode, welche den mittleren quadratischen logarithmischen Fehler (MSLE) minimiert deutlich bessere Ergebnisse erzielt als bisher bekannte Verfahren. Hierbei wird die Güte der Parameterschätzung, die Genauigkeit der PSD-Schätzung sowie das Verhalten des Gesamtsystems hinsichtlich der Unterdrückung der Restechos zum Vergleich herangezogen.

Drittens wird ein neuartiges Modell für das PSD des frühen Restechos vorgeschlagen und mit dem IIR-Filter-Modell für das PSD des späten Restechos kombiniert. Dabei entsteht ein neuartiges Modell für das PSD des gesamten Restechos. Hierbei wird das PSD des frühen Restechos durch einen zeitlich gleitenden Mittelwert auf dem PSD der Lausprechersignale dargestellt welcher durch einen Kopplungsfaktor auf das zu schätzende PSD des frühen Restechos abgebildet wird. Weiter werden signalbasierte Methoden vorgeschlagen welche auf der Methode des Ausgangsfehlers beruhen und nun alle drei Parameter des neuartigen Gesamtmodells gemeinsam schätzen. Auch hier wird als Kostenfunktion der MSLE minimiert. Die Ergebnisse der Simulationen zeigen, dass die Ausgangsfehlermethode, hinsichtlich Güte der Parameterschätzung sowie der Genauigkeit der PSD-Schätzung auch hier deutlich bessere Ergebnisse erzielt als vergleichbare bereits bekannte Verfahren. Im Vergleich zu den momentan bekannten RES-Methoden liefert die vorgeschlagene Methode den besten Wert für das segmentelle „Speech-to-Speech-Distortion Ratio“ (ca. 2-5 dB besser) und erreicht gleichzeitig den besten Wert für die segmentelle Restecho-Dämpfung (ca. 1-2 dB besser).

GLOSSARY

Acronyms and abbreviations

3DM	three-dimensional M-Max
AEC	acoustic echo cancellation
AP	affine projection
cXM	center-clipping exclusive-maximum
DEA	dynamic effort allocation
DNN	deep neural network
EE	equation error
ERE	early residual echo
ERLE	echo return loss enhancement
FEA	fixed effort allocation
FFT	fast Fourier transform
FIR	finite impulse response
FLMS	fast least mean squares
IIPNLMS	improved IPNLMS
IIR	infinite impulse response
IPMDF	improved proportionate multidelay filtering
IPNLMS	improved proportionate normalized least mean squares
IR	impulse response
LEM	loudspeaker-enclosure-microphone
LRE	late residual echo
LSD	log spectral distance
MAEC	multichannel acoustic echo cancellation
MDF	multidelay filtering
MSE	mean squared error
MSLE	mean squared log error
NEC	network echo cancellation
NL	non-linear
NLMS	normalized least mean squares
OE	output error

PB-FDAF	partitioned block frequency-domain adaptive filtering
PLR	pseudo-linear regression
PNLMS	proportionate normalized least mean squares
PSD	power spectral density
PUNLMS	partial update normalized least mean squares
REA	residual echo attenuation
RES	residual echo suppression
RIR	room impulse response
RLS	recursive least squares
RPE	recursive prediction error
SC-IPMDF	sparseness-controlled IPMDF
SC-IPNLMS	sparseness-controlled IPNLMS
SPU	selective-partial-update
SPUNLMS	selective-partial-update normalized least mean squares
SRER	speech-to-residual echo ratio
SSDR	speech-to-speech distortion ratio
STFT	short-time Fourier transform
UFLMS	unconstrained fast least mean squares
WOLA	weighted overlap-add
XM	exclusive-maximum

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Hands-free speech communication devices have seen a steady rise in their availability and use during the last few decades. Their popularity has been driven primarily due to their ease-of-use, as speech is a highly natural and effective means of communication, coupled with significant improvements in hardware technology and signal processing algorithms. Hands-free speech communication devices have been deployed for a wide variety of applications, such as teleconferencing, in-car communication, assisted learning and as personal assistants. In particular, speech-enabled consumer electronic devices, such as voice-controlled televisions, smartphones and portable smartspeakers, have seen a huge boost in sales in recent years due to the ubiquitousness of automatic speech recognition technology. They are used in a diverse range of physical and acoustic environments, which require them to cope with specific and increasingly complex demands imposed by users, acoustic environments and applications.

Speech-enabled communication devices are typically equipped with one or more microphones and one or more loudspeakers. In the context of telephony the loudspeakers play back the voice of the person on the other end of the call, while in the context of voice-controlled devices the loudspeakers play back, e.g., a speech prompt or music. In addition to capturing the desired speech from the user, the microphones also pick up acoustic echo due to the acoustic coupling between the loudspeakers and the microphones, and other undesired interferences such as background noise [1–5]. At increased levels, acoustic echo and background noise may degrade speech quality, speech intelligibility and the performance of automatic speech recognition systems [5–8].

Acoustic echo control has been a popular area of research [1–4, 9–14], where systems aiming at eliminating acoustic echo typically employ a combination of acoustic echo cancellation and residual echo suppression techniques. An acoustic echo cancellation (AEC) system employs an adaptive filter to compensate for the acoustic echo path between the loudspeakers and microphones. When deploying such a system in a highly reverberant environment, a very long AEC filter with several thousand taps is required to accurately model the echo path and effectively cancel the acoustic echo. Using such a long filter naturally results in large computational cost for updating the filter coefficients and may also lead to slow filter convergence [1, 10, 15].

In order to reduce the computational complexity of the filter update, several approaches have been proposed. A popular approach is to reduce the effort for updating the AEC filter by focusing only on the most important filter taps. Several so-called tap selection schemes have been proposed for implementing such partial updates of the AEC filter [16–20]. A more straight-forward approach for reducing computational complexity is to reduce the length of the AEC filter. This, however, has the undesirable effect of the AEC filter being incapable of modeling the complete acoustic echo path, possibly leading to a significant amount of residual echo. Many residual echo suppression postfilters have been proposed, mostly implemented as a spectral weighting in the subband domain [1,13,14,21,22]. To compute the spectral weights, an accurate estimate of the power spectral density (PSD) of the residual echo is required, for which several model-based methods have been proposed [13,14,23,24].

The main objectives of this thesis are to **investigate and develop tap selection schemes for multichannel systems** in order to achieve low-complexity acoustic echo cancellation, and to **improve model-based residual echo PSD estimators** in order to achieve effective residual echo suppression.

In this chapter, we present a general introduction to the problem, state-of-the art solutions and an outline of the thesis. Section 1.1 presents the typical acoustic scenario in which hands-free speech communication devices are deployed. Section 1.2 provides an overview of different techniques that are commonly used to achieve low-complexity acoustic echo cancellation. Section 1.3 provides an overview of state-of-the-art residual echo suppression techniques. Section 1.4 presents the main contributions and a chapter-by-chapter overview of the thesis.

1.1 Acoustic scenario

Fig. 1.1 illustrates a typical acoustic scenario in which hands-free speech communication devices are often used. The device is placed inside an enclosure, such as a room or a car, and the user can interact with the device using their voice. Based on the application, environment and the physical design of the device, the desired speech from the user may become heavily corrupted by multiple undesired interfering sources. This presents many challenges which need to be tackled in order to achieve high-quality speech communication.

A major source of interference in such a scenario is the acoustic echo, which occurs when the microphone picks up sound radiated by the loudspeaker and its reflections against the borders of the enclosure as well as against other objects [1,2]. In the context of telephony, the person on the other end of the call, referred to as the far-end speaker, may become annoyed by listening to their own voice with a certain delay, thereby hampering communication. For voice-controlled devices, acoustic echo may cause the speech recognizer to transcribe the user commands incorrectly.

The acoustic echo path between the loudspeaker and the microphone inside this loudspeaker-enclosure-microphone (LEM) system can be described by an impulse

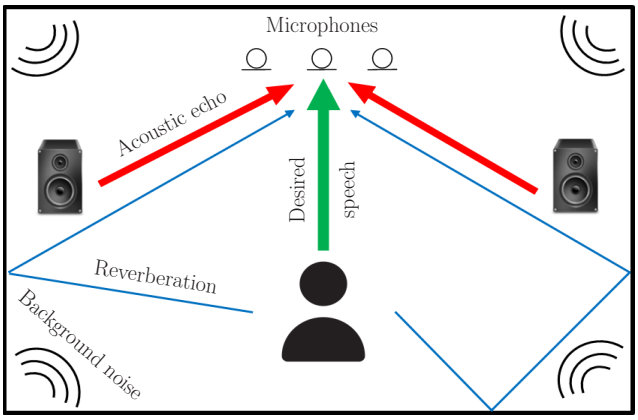

Fig. 1.1: A typical acoustic scenario in which a hands-free speech communication device is used, with desired speech from the user and undesired interferences such as acoustic echo, reverberation and background noise.

response (IR). Fig. 1.2 shows a typical IR of an LEM system, which is characterized by the following three components [25]:

- *direct path* - initial peak corresponding to the direct sound, arriving at the microphone at a delay depending on the distance between the loudspeaker and the microphone,
- *early reflections* - distinct impulses with relatively large amplitudes, dependent on the geometry of the enclosure and the positioning of the device within the enclosure, and
- *late reflections* - densely-spaced impulses with relatively small amplitudes, decaying in nature; also known as the reverberant part.

An important room acoustical metric is the reverberation time (T_{60}), which is the time it takes for sound energy to decay by 60 dB compared to the direct sound component [25, 26]. The reverberation time typically depends on the geometry of the enclosure and the reflectivity of the surfaces [27], with larger enclosures typically characterized by larger reverberation times. In [25] it has been shown that the T_{60} is generally frequency-dependent. A statistical reverberation model for an IR was proposed in [28], which describes the late reverberant part of an IR as a realization of a stochastic process that decays exponentially at a rate proportional to the T_{60}. In [26], statistical methods were introduced to calculate the reverberation time of an enclosure irrespective of its geometry, while in [29], the T_{60} was estimated in a frequency-independent manner via line-fitting on the energy decay curve of the IR.

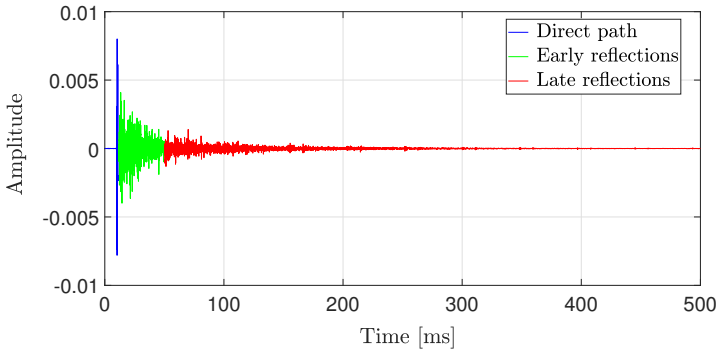

Fig. 1.2: An example of an impulse response (reverberation time $T_{60} \approx 500$ ms)

1.2 Low-complexity acoustic echo cancellation

An AEC system employs an adaptive filter to compensate for the acoustic echo path between the loudspeaker and the microphone [1,2,9,10]. Fig. 1.3 illustrates an AEC system, consisting of an adaptive finite impulse response (FIR) filter, which takes the loudspeaker signal as input, and generates an estimate of the acoustic echo signal. This estimated echo signal is subtracted from the microphone signal, with the resulting signal referred to as the AEC error signal. The filter coefficients are updated in order to minimize a cost function, e.g., the energy of the error signal. From a signal processing perspective, this can be seen as a system identification problem, where the filter aims at adaptively estimating the IR of the LEM system. For a highly reverberant environment, with a T_{60} of several hundred milliseconds, this would necessitate the use of a long FIR filter with several thousand filter taps in order to effectively reduce the acoustic echo. This would result in large computational cost and large memory requirements for the filter update, and may also lead to slow filter convergence [1,10,15]. Thus, a trade-off exists between computational cost and echo cancellation performance, such that the AEC filter length should be carefully chosen for AEC systems in challenging acoustic scenarios.

In order to update the adaptive filter, many algorithms such as the normalized least mean squares (NLMS), affine projection (AP) and recursive least squares (RLS) have been proposed [1, 15]. These algorithms differ in terms of the cost function minimized, their filter convergence and tracking behaviour, as well as their computational complexity. On the one hand, the NLMS algorithm delivers the slowest convergence and tracking performance among these algorithms, with the RLS algorithm delivering the fastest initial convergence and the AP algorithm delivering the fastest tracking performance, respectively [10]. On the other hand, the NLMS algorithm has the lowest computational complexity and memory requirements among these algorithms, while the AP and RLS algorithms have higher computational complexity as well as large memory requirements [1].

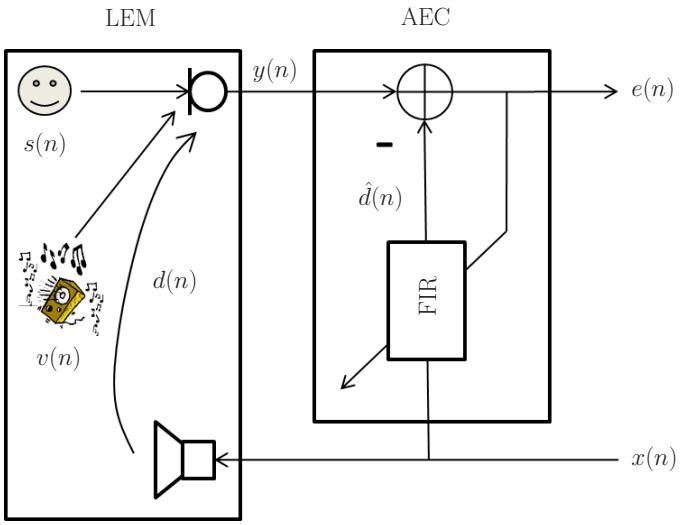

Fig. 1.3: Schematic illustration of typical LEM and AEC systems, where $x(n)$, $s(n)$, $v(n)$, $d(n)$, $y(n)$, $\hat{d}(n)$, and $e(n)$ denote the loudspeaker signal, desired speech, background noise, acoustic echo, microphone signal, estimated echo signal and the AEC error signal, respectively.

As the aim of this thesis is to achieve low-complexity AEC, we will only consider the NLMS algorithm for updating the adaptive filter. In order to further reduce the computational complexity, in Section 1.2.1 we give an overview of adaptive algorithms that employ tap selection schemes for partially updating the AEC filter. Section 1.2.2 presents an overview of frequency and subband-domain adaptive algorithms which are often used to further reduce computational cost and to enable frequency-selective filter updates.

1.2.1 *Partial-update adaptive filtering using tap selection schemes*

In order to reduce the computational complexity of the AEC filter update, a number of partial-update adaptive filtering algorithms have been proposed [16–20, 30–32], which save computations by updating only a fixed subset M of all N filter coefficients in each iteration. Most of these algorithms are variants of the NLMS algorithm and use tap selection schemes based on different criteria to determine which filter taps should be selected for update in each iteration. In Section 1.2.1.1, we discuss tap selection schemes based on predefined schedules. In Sections 1.2.1.2 and 1.2.1.3, we discuss tap selection schemes which exploit the sparsity present in the input signals and the echo path, respectively, while in Section 1.2.1.4, we discuss tap selection schemes for multichannel AEC systems.

1.2.1.1 *Schemes based on predefined schedules*

A simple criterion for performing tap selection is to use a predefined scheduling scheme. Notable examples are the sequential-block NLMS algorithm [16], which updates the filter coefficients sequentially by updating a continuous block of M coefficients in each iteration, and the sequential NLMS algorithm [17], which updates the filter coefficients periodically by updating each coefficient only once every $\frac{N}{M}$ iterations. These algorithms have almost no computational overhead but suffer from slow convergence as more iterations are needed to update all filter coefficients an equal number of times.

1.2.1.2 *Schemes based on exploiting the sparsity of the input signal*

In contrast to predefined scheduling schemes, most partial-update algorithms use tap selection schemes which exploit some underlying information about the acoustic environment, e.g., the sparsity present in the input loudspeaker signal. A signal can be considered sparse if a large fraction of its energy is concentrated in a small fraction of its samples. Since the loudspeaker signal is typically a speech or music signal, it usually exhibits significant sparsity across frequency, due to spectrally coloured content, and across time, due to non-stationarity. One of the first proposed algorithms which performs tap selection by exploiting input signal sparsity is the max-NLMS algorithm [30], which updates the filter coefficient with the largest magnitude tap-input, i.e., only 1 out of N filter coefficients is updated in each iteration. The M-Max NLMS algorithm [18] extends this concept by updating the filter coefficients with the M largest magnitude tap-inputs in each iteration. For a given M, this algorithm yields the closest possible performance compared to full-update NLMS in terms of minimizing the squared a-posteriori error [18]. The selective-block-update NLMS [32] and the selective-partial-update NLMS (SPUNLMS) [19] algorithms extend the M-Max NLMS algorithm to coefficient blocks in order to reduce memory requirements. The SPUNLMS algorithm divides the N-tap adaptive filter into U equal-sized blocks, ranks them according to the squared Euclidean norm of their tap-inputs and updates all filter coefficients in the top $U \cdot \frac{M}{N}$ ranked blocks in each iteration. Compared to predefined scheduling schemes, these tap selection schemes have higher computational overhead but generally outperform them in terms of convergence speed and AEC performance. Compared to the full-update NLMS algorithm, they have lower computational complexity but usually suffer from decreased convergence speed.

1.2.1.3 *Schemes based on exploiting the sparsity of the impulse response*

It was shown in [33] that using the full-update NLMS algorithm gives an unsatisfactory performance when the echo path is sparse. Therefore, a whole family of partial-update adaptive filtering algorithms have been proposed for applications with sparse IRs, such as network echo cancellation (NEC) [34–40]. When an IR is sparse, a small

number of its samples contain the majority of its energy, and therefore contribute the most to the echo signal. The proportionate NLMS (PNLMS) algorithm [35] exploits the sparsity of the IR to perform tap selection by updating each filter coefficient with a step-size proportionate to its magnitude. This enables the larger coefficients to reach their optimal values in fewer number of iterations, resulting in a faster initial reduction of the echo. The PNLMS algorithm has been shown to outperform algorithms which use uniform step-sizes for all coefficients, such as the NLMS algorithm, for sparse system identification but performs poorly for non-sparse and/or time-varying systems. In order to tackle these problems, improved versions of the PNLMS algorithm have been proposed. The improved PNLMS (IPNLMS) algorithm [36] updates the coefficients using a mix of the NLMS and PNLMS algorithm in every iteration with the help of a single weighting factor. The sparseness-controlled IPNLMS (SC-IPNLMS) algorithm [40] incorporates a sparsity metric to compute this weighting factor. The improved IPNLMS (IIPNLMS) algorithm [37] enables the amount of proportionate and non-proportionate update of each coefficient to be controlled independently of each other, i.e., a different weighting factor for each coefficient. Finally, the sparse partial-update NLMS algorithm [38] performs tap selection by exploiting the sparsity of both the tap-inputs as well as the echo path. In each iteration, it selects those M filter taps for update which yield the largest magnitude outputs.

Since acoustic IRs are not particularly sparse, in this thesis we will not consider any tap selection schemes which exploit the sparsity of the echo path and instead only focus on tap selection schemes which exploit input signal sparsity.

1.2.1.4 *Schemes for multichannel acoustic echo cancellation*

Many modern voice-controlled entertainment devices, such as surround-sound systems and smart TVs, are equipped with multiple microphones and loudspeakers in order to deliver a higher-quality experience to the users. In order to compensate for the acoustic coupling between each loudspeaker and each microphone, such devices need to employ multichannel acoustic echo cancellation (MAEC), with a separate adaptive FIR filter used to estimate the IR between each loudspeaker-microphone pair. This drives up the computational load by a multiplicative factor when using the full-update NLMS algorithm.

For most applications, the loudspeaker signals are heavily correlated with each other. For example, in the context of teleconferencing, the signals captured by the microphones on the transmitting device are filtered versions of the same sound source (the far-end speaker). It has been shown that for such systems a non-uniqueness problem exists [41, 42], i.e., the adaptive filters are not able to uniquely identify the IRs between the loudspeakers and microphones (of the receiving device). This by itself would not be a major problem as long as the acoustic conditions in the transmission room are fixed, as the multiple adaptive filters would nevertheless work together to minimize the acoustic echo. However, it has been shown that an abrupt change in the characteristics of the loudspeaker signals, e.g., due to a change in the position

of the transmitting device with respect to the far-end speaker, would require the filters to start the adaptation process afresh.

Even though the non-uniqueness problem is somewhat softened in practice due to the so-called tail effect [42], it has been shown that the multichannel tap-input covariance matrix is highly ill-conditioned due to the high coherence between the input signals [42, 43], resulting in large filter misalignment and slow convergence speed [41–43]. To tackle this challenging problem, several techniques have been proposed, aiming at decorrelating the input signals while not affecting their audio quality. Potential solutions include using interleaving comb filters [41], adding spectrally-shaped random noise to the input signals [44], modulating the input signals [45], using time-varying all-pass filters [46] and non-linear processing of the input signals [2, 42]. In addition, tap selection schemes have been proposed to specifically tackle the misalignment problem for stereo (2-channel) AEC systems such as the exclusive-maximum (XM) scheme [47–49] and the center-clipping XM (cXM) scheme [50]. The XM scheme improves the condition number of the input covariance matrix by prohibiting the selection of the same filter tap index in both filters for update, and at the same time maximizes the squared Euclidean norm of the selected tap-inputs in each iteration. The XM tap selection scheme was combined with non-linear processing of the input signals to yield the XMNL-NLMS algorithm [47–49]. The cXMNL-NLMS algorithm [50] further improves the convergence speed by employing a center-clipping algorithm, which makes it more robust to the positioning of the transmitting device with respect to the far-end speaker.

Even though tap selection schemes have been shown to alleviate the misalignment problem for MAEC systems, our main motivation for developing tap selection schemes is to reduce the computational complexity compared to full update algorithms.

1.2.2 *Frequency-domain and subband-domain processing*

In order to further reduce computational cost and allow for frequency-selective filter updates, frequency-domain and subband adaptive filters have been proposed [51,52]. A comprehensive comparison of the advantages and disadvantages between time-domain, frequency-domain and subband processing techniques in terms of computational complexity, performance and signal delay has been provided in [1].

Frequency-domain adaptive filtering algorithms such as fast-LMS (FLMS) [53] and unconstrained FLMS (UFLMS) [54] incorporate a block updating strategy, where the input block length is typically chosen equal to the filter length N. The convolution, filtering and adaptation operations are performed in the frequency domain using the efficient fast Fourier transform (FFT) [55], where the output is computed once every N samples using the overlap-save method [15,51], thereby reducing computational cost. This however introduces a delay of N samples between the input and output signals, which may not be desirable for some applications. Therefore, the multidelay filtering (MDF) algorithm was proposed in [56], where the filter is

partitioned into U equal-sized blocks, thereby reducing the delay by a factor of U compared to the FLMS algorithm. Tap selection schemes such as the M-Max and SP schemes have been incorporated into the MDF algorithm to further reduce complexity and improve convergence for NEC applications [57]. Additionally, algorithms which exploit the sparsity of the echo path such as IPMDF [58] and SC-IPMDF [59] have been proposed based on the MDF algorithm for NEC and AEC applications, respectively.

In subband adaptive algorithms, the input (loudspeaker) signal and the microphone signal are split into multiple frequency subbands using an analysis filterbank [60], e.g., using the short-time Fourier transform (STFT). In each subband, the echo estimate is generated via convolution of the input signal and the subband adaptive filter, which is then subtracted from the microphone signal to yield the subband error signal. The filter adaptation is performed independently in each subband, which allows for frequency-dependent step-size control. A synthesis filterbank is then used to reconstruct the error signal back into the time domain using the weighted overlap-add (WOLA) method [1, 61].

In this thesis, subband processing using an FFT-based analysis filterbank and inverse FFT synthesis filterbank is considered, and the subband adaptive filters used for MAEC are updated using the subband NLMS algorithm. In order to further reduce computational cost, different tap selection schemes are implemented in the subband domain which exploit the sparsity present in the multichannel input signals across frequency, channels and time.

1.3 Residual echo suppression

In practice, the AEC filter is typically unable to perfectly estimate the echo path, resulting in residual echo due to the mismatch between the true and estimated echo paths [1, 13]. As the main energy content of the IR of an LEM system is typically concentrated in the first 50-100 ms [25,27], a short AEC filter is often used in practice in order to save computational cost and increase convergence speed. However, this under-modeling of the echo path may result in a significant amount of late residual echo, especially in reverberant environments. Residual echo may also occur due to sudden changes in the echo path, or due to non-linear echo components caused by low-cost hardware.

In order to achieve high-quality speech communication, it is necessary to effectively suppress the residual echo. Residual echo suppression (RES) is typically performed by means of a subband-domain postfilter [1,13,14,21,22,62–67], as illustrated in Fig. 1.4. Although multi-frame postfilters have been proposed [68], most postfilters apply a real-valued weight in each subband to attenuate the residual echo. In addition to the residual echo, the postfilter may also be used to suppress other interferences such as reverberation [14] and background noise [21,22]. In addition to the postfilter, non-linear post-processing may further be applied [1]. The RES postfilter typically requires an estimate for the PSD of the residual echo in order to compute the

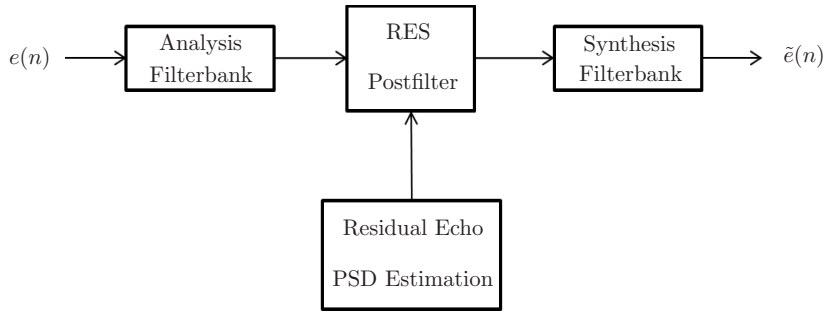

Fig. 1.4: Schematic illustration of a typical residual echo suppression (RES) system, where $e(n)$ and $\tilde{e}(n)$ denote the AEC error signal and its postfiltered version, respectively.

spectral weights for each subband, which makes it necessary to accurately estimate the residual echo PSD. In this thesis, we will only consider residual echo caused by linear echo components, which can further be divided into early residual echo due to the mismatch between the true and the estimated echo paths (filter misalignment), and late residual echo due to under-modeling of the echo path by a short AEC filter. In Sections 1.3.1 and 1.3.2, we present models and estimators for the PSD of the early and late residual echo components, respectively.

1.3.1 *Residual echo due to filter misalignment*

In [1], it is shown that the early residual echo PSD is related to the PSD of the loudspeaker signal as well as the amount of filter misalignment. It is proposed to model the relationship between the PSD of the loudspeaker signal and the early residual echo PSD using a single-tap real-valued frequency-dependent coupling factor, where the coupling factor represents the squared magnitude of the filter misalignment in the subband domain. In [64], it is proposed to estimate the early residual echo PSD by applying a coupling factor to the PSD of the estimated echo signal (instead of the PSD of the loudspeaker signal), while in [65] two coupling factors are used to separately estimate the PSD of the early residual echo due to linear and non-linear echo components, respectively. In [69], a pure acoustic echo suppression system is considered, i.e., without an AEC filter, and a coupling factor is applied to the PSD of the loudspeaker signal to directly estimate the acoustic echo PSD. In each of these cases, the coupling factor is estimated only during periods of local speech absence, i.e., when the microphone signal is dominated by the residual echo, by computing a smoothed ratio between the PSD of the microphone signal and the PSD of the loudspeaker signal (or the PSD of the estimated echo signal). A potential drawback of using an estimator based on a model with a single parameter, such as a single-tap coupling factor, is that it may not be completely successful in estimating the entire residual echo due to filter misalignment, as the misalignment may be spread over all taps of the AEC filter [1, 70, 71].

1.3.2 *Residual echo due to under-modeling of the echo path*

Several estimators have been proposed to estimate the PSD of the residual echo that occurs due to the deficient length of the AEC filter [13, 14, 23]. These estimators assume that the AEC filter is able to compensate for the direct path and early reflections, but is not long enough to model the late reflections, thereby leading to late residual echo. A similar estimator has also been proposed in the context of acoustic echo suppression (i.e., without an AEC filter) [24]. The estimators in [13, 14, 24] are based on the statistical reverberation model for an IR proposed in [28], which describes the late reverberant part of an IR as a realization of a stochastic process that decays exponentially at a rate proportional to the T_{60}, while the estimator in [23] is based on a similar statistical reverberation model for an IR proposed in [72]. All estimators compute the PSD of the late residual echo recursively and rely on two model parameters: the reverberation scaling parameter, which is related to the initial power of the late residual echo, and the reverberation decay parameter, which is related to the T_{60} of the IR. Both channel-based as well as signal-based methods have been proposed to estimate these two model parameters.

Channel-based methods estimate the parameters from the estimated echo path, i.e., the coefficients of the converged AEC filter [13, 14]. In [13], the reverberation parameters are assumed to be frequency-independent and are estimated by applying a direct fit to the log-envelope of the AEC filter coefficients. In [14], the reverberation parameters are assumed to be frequency-dependent and are estimated in online mode. First, a linear curve is fitted in each subband to an appropriately selected portion of the energy decay curve of the AEC filter coefficients. The slope of this line is then used to estimate the T_{60} and the reverberation decay parameter. The reverberation scaling parameter is eventually estimated by extrapolating the energy of the AEC filter coefficients using the estimated reverberation decay parameter.

In contrast to channel-based methods, signal-based methods estimate the reverberation parameters directly from the loudspeaker and residual echo signals. In [23], the parameters are assumed to be frequency-dependent and estimated in offline mode (i.e., batch processing). First, the reverberation decay parameter is estimated by minimizing the mean squared log error (MSLE) between the AEC error PSD and the estimated late residual echo PSD. The estimated reverberation decay parameter is then used to estimate the reverberation scaling parameter by minimizing the mean squared error (MSE) between the AEC error PSD and the estimated late residual echo PSD. In [24], an online method exploiting higher-order-statistics has been proposed to estimate the frequency-dependent parameters independently of each other.

It is important to note that channel-based methods work better when longer AEC filters are used, as the AEC filter coefficients are more likely to successfully capture the decay of the late reverberant part of the IR. Although signal-based methods can in principle be used for any AEC filter length, the AEC filter should still be long enough to compensate for the direct path and early reflections, such that the assumed reverberation model is valid. To the best of our knowledge, no signal-based

methods have been proposed that jointly estimate both reverberation parameters by minimizing a single cost function in online mode.

1.4 Outline of the thesis and main contributions

The main objectives of this thesis are to develop and evaluate low-complexity adaptive algorithms for multichannel acoustic echo cancellation (MAEC) and model-based algorithms for residual echo suppression (RES). Throughout the thesis, we consider subband processing using an FFT-based analysis filterbank, where we use subband adaptive filters for MAEC which are updated using the subband NLMS algorithm. Aiming at reducing the computational complexity of the filter update in MAEC systems, we focus on tap selection schemes which exploit the sparsity in the multichannel input signals. Aiming at improving the RES performance, we focus on model-based residual echo PSD estimators based on joint estimation of all model parameters.

The main contributions of this thesis can be summarized as follows. First, **we propose novel tap selection schemes which exploit input signal sparsity across the dimensions of frequency, channels and time**, leading to efficient partial updates of adaptive MAEC filters in the subband domain. In particular, the proposed dynamic effort allocation scheme proportionately selects more filter taps for update in subbands and channels with larger magnitude tap-inputs while not ignoring the filters with smaller magnitude tap-inputs. Simulation results show that the proposed tap selection scheme achieves almost identical echo cancellation performance compared to updating all filter taps at a significantly reduced computational cost. Second, **we propose novel signal-based methods to estimate the late residual echo PSD in online mode**. The late residual echo PSD is modeled using an infinite impulse response (IIR) filter on the PSD of the loudspeaker signal, based on frequency-dependent reverberation scaling and decay parameters. We propose several signal-based methods based on output error and equation error to jointly estimate both reverberation parameters by minimizing a single cost function in online mode. Simulation results show that the output error method minimizing the MSLE cost function outperforms state-of-the-art offline and online methods in terms of parameter estimation accuracy, late residual echo PSD estimation accuracy and residual echo suppression performance. Third, **we propose a novel model combining early and late residual echo PSDs and propose signal-based methods to jointly estimate all required model parameters in online mode**. We propose to model the early residual echo PSD using a moving average filter on the PSD of the loudspeaker signal, based on a frequency-dependent coupling factor. The late residual echo PSD is again modeled using an IIR filter based on frequency-dependent reverberation scaling and decay parameters. We propose signal-based methods based on output error to jointly estimate all three model parameters, i.e., the coupling factor and the reverberation scaling and decay parameters, by minimizing a single MSLE cost function in online mode. Simulation results show that the proposed output error method with the recursive prediction error algorithm outperforms state-of-the-art offline and online methods in terms of

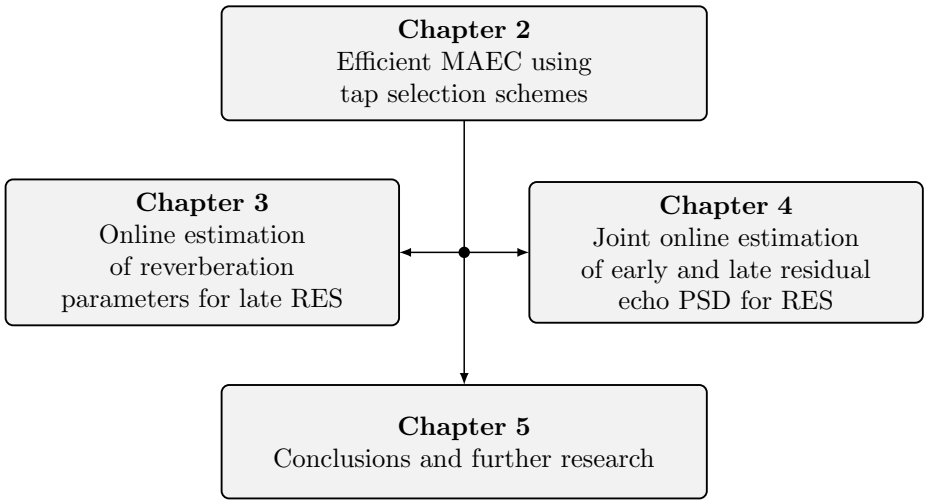

Fig. 1.5: Structure of the thesis.

residual echo PSD estimation accuracy, maximizing residual echo suppression and minimizing speech distortion.

In the remainder of this section, we provide a chapter-by-chapter overview of this thesis, summarizing the content and the contributions of each chapter. A schematic overview of the thesis is given in Fig. 1.5.

In **Chapter 2**, we start by analyzing the sparsity present in real-world multichannel signals across the three dimensions of frequency, channels and time. Based on this analysis, we develop tap selection schemes which exploit this sparsity for partially updating the adaptive filters of MAEC systems. We first investigate the three-dimensional M-Max (3DM) scheme, which is an extension of the M-Max scheme for the single-channel scenario to the multichannel scenario, and the SPU scheme, which only selects taps in filters with the largest magnitude tap-inputs. Since these tap selection schemes are based on the M-Max criterion, they completely ignore the filters with the smallest magnitude tap-inputs for update. In order to overcome this problem, we propose two novel tap selection schemes which do not ignore any filters for update. On the one hand, the fixed effort allocation (FEA) scheme selects a fixed number of filter taps in each subband and channel, thereby not exploiting signal sparsity across frequency and channels. On the other hand, the dynamic effort allocation (DEA) scheme exploits signal sparsity by dynamically allocating filter taps in a proportionate manner, i.e., more taps are selected in filters with relatively larger magnitude tap-inputs while the filters with the smallest magnitude tap-inputs are not completely ignored. Simulation results for speech and multichannel music signals show that the 3DM scheme and the proposed DEA scheme deliver the best echo cancellation performance. These schemes achieve almost identical echo return loss enhancement scores compared to full filter update (about 1 dB worse), even when only 20% of all filter taps are updated in every frame. However, the 3DM scheme

still requires about 94% of the total computational cost needed for full filter update when updating only 20% of all filter taps, while the proposed DEA scheme requires only about 28%, thereby significantly saving computational cost. This chapter has been published as a journal paper in the EURASIP Journal on Advances in Signal Processing [73]. Another publication related to this chapter is [74].

In **Chapter 3**, we model the relationship between the PSD of the late residual echo (caused by under-modeling of the echo path by the AEC filter) and the PSD of the loudspeaker signal using an IIR filter with two frequency-dependent coefficients, namely the reverberation scaling and decay parameters. We propose two signal-based methods (output error and equation error) to jointly estimate both coefficients of the IIR filter by minimizing a single MSE or MSLE cost function in online mode. These signal-based methods were originally proposed to estimate the coefficients of time-domain IIR filters and we apply them to PSDs to estimate the reverberation parameters. For the output error method, we use gradient-descent-based algorithms such as recursive prediction error (RPE) and pseudo-linear regression (PLR) [75] to jointly estimate both reverberation parameters. The proposed parameter estimation methods are first evaluated in an idealistic setting using artificially generated IRs, no filter misalignment, no near-end speech and no background noise signals. Simulation results show that the proposed output error method with the RPE algorithm when minimizing the MSLE cost function outperforms all other proposed methods in terms of estimation accuracy of the reverberation parameters as well as the late residual echo PSD. This result is similar to the results obtained for offline processing. The proposed methods are then compared with state-of-the-art offline and online parameter estimation methods in a realistic setting using IRs measured in different rooms, a fully converged subband AEC filter, near-end speech and background noise signals. Simulation results show that the proposed output error method with the RPE algorithm when minimizing the MSLE cost function outperforms state-of-the-art methods, as well as all other proposed methods, in terms of estimation accuracy of the reverberation time and the late residual echo PSD. Additionally, it yields the best segmental speech-to-speech distortion ratio score (about 5-10 dB better than most methods), while not losing too much in terms of the segmental residual echo attenuation score (about 2-3 dB worse than most methods). This chapter has been published as a journal paper in the IEEE/ACM Transactions on Audio, Speech and Language Processing [78]. Another publication related to this chapter is [77].

Building upon the methods proposed in **Chapter 3** for late residual echo suppression, in **Chapter 4** we consider both late residual echo as well as early residual echo (caused by filter misalignment). We propose to model the relationship between the PSD of the early residual echo and the PSD of the loudspeaker signal using a moving average filter with the same filter length as the subband AEC filter. Assuming that the misalignment is spread evenly over all AEC filter taps, we propose to model the coefficients of the moving average filter using a frequency-dependent coupling factor. This moving average filter model for the early residual echo PSD can be interpreted as extending the single-tap frequency-dependent coupling factor model over all taps of the AEC filter. The proposed moving average filter model for the early residual echo PSD is combined with the IIR filter model for the late residual

echo PSD in Chapter 3 to yield a new model for the residual echo PSD. We propose signal-based methods to jointly estimate all three model parameters (both reverberation parameters and the coupling factor) in online mode. In particular, based on the results obtained in Chapter 3, we use the output error method with the RPE and PLR algorithms and minimize a single MSLE cost function. The proposed methods estimating the three model parameters (3P) are extensions of the methods in Chapter 3, which only estimate both reverberation parameters (2P). The proposed 3P methods are first compared with their simplified 2P versions in an idealistic setting using artificially generated IRs, no near-end speech and no background noise signals. Simulation results show that the proposed 3P methods yield accurate estimates for all three model parameters as well as the residual echo PSD, irrespective of the amount of filter misalignment, with the RPE algorithm performing better than the PLR algorithm, while the 2P methods fail completely when high amounts of filter misalignment are present. The proposed 3P methods are then compared with state-of-the-art offline and online parameter estimation methods in a realistic setting using IRs measured in different rooms, a pre-converged subband AEC filter, near-end speech and background noise signals. Simulation results show that the proposed 3P method with the RPE algorithm outperforms all considered methods in terms of estimation accuracy of the residual echo PSD and yields the best segmental speech-to-speech distortion ratio score (about 2-5 dB better than other methods), while also yielding the best segmental residual echo attenuation score (about 1-2 dB better than other methods). This chapter has been submitted as a journal paper to the IEEE/ACM Transactions on Audio, Speech and Language Processing [79].

In **Chapter 5**, we summarize the main contributions of the thesis and present suggestions for further research.

2

EFFICIENT MULTICHANNEL ACOUSTIC ECHO CANCELLATION USING CONSTRAINED TAP SELECTION SCHEMES IN THE SUBBAND DOMAIN

2.1 Abstract

Acoustic echo cancellation (AEC) is a key speech enhancement technology in speech communication and voice-enabled devices. AEC systems employ adaptive filters to estimate the acoustic echo paths between the loudspeakers and the microphone(s). In applications involving surround sound, the computational complexity of an AEC system may become demanding due to the multiple loudspeaker channels and the necessity of using long filters in reverberant environments. In order to reduce the computational complexity, the approach of partially updating the AEC filters is considered in this paper. In particular, we investigate tap selection schemes which exploit the sparsity present in the loudspeaker channels for partially updating subband AEC filters. The potential for exploiting signal sparsity across three dimensions, namely time, frequency and channels, is analyzed. A thorough analysis of different state-of-the-art tap selection schemes is performed and insights about their limitations are gained. A novel tap selection scheme is proposed which overcomes these limitations by exploiting signal sparsity while not ignoring any filters for update in the different subbands and channels. Extensive simulation results using both artificial as well as real-world multichannel signals show that the proposed tap selection scheme outperforms state-of-the-art tap selection schemes in terms of echo cancellation performance. In addition, it yields almost identical echo cancellation performance as compared to updating all filter taps at a significantly reduced computational cost.

2.2 Introduction

Acoustic echo cancellation (AEC) [1, 2] is a key technology used in hands-free telephony and voice-enabled systems. An AEC system consists of an adaptive filter which estimates the acoustic echo path between the loudspeaker and the micro-

phone. Using this estimated echo path, an estimate of the acoustic echo signal is generated which is then subtracted from the microphone signal. When multiple loudspeakers are present, as is the case for surround-sound systems, Multichannel Acoustic Echo Cancellation (MAEC) systems are required [41, 80–82]. These systems consist of multiple adaptive filters dedicated to estimate the acoustic echo paths between each loudspeaker and each microphone, i.e., one filter per channel. When employing time-domain MAEC systems in large and/or reverberant rooms, very long filters with several thousand taps may be required in order to achieve effective echo cancellation. Using such long filters requires large computational effort, both for updating the filters as well as for generating the acoustic echo signal estimates.

In order to reduce computational complexity of time-domain adaptive filters, a number of tap selection schemes [17–20, 31, 32, 38, 83] have been proposed for implementing partial updates of the adaptive filters. These schemes reduce complexity by updating only a subset M of all N filter taps in each iteration, where the subset is chosen based on a tap selection criterion. Since speech and/or surround-sound entertainment signals usually exhibit significant sparsity across frequency (due to spectrally colored content), channels (due to different content in the different loudspeakers) and time (due to non-stationary content), a number of tap selection schemes have been proposed which exploit the sparsity present in the loudspeaker signals for partially updating the filters [18–20, 31, 32]. The M-Max [18, 31] is a well-known tap selection scheme which exploits signal sparsity by selecting the filter taps corresponding to the M largest magnitude tap-inputs in each iteration. For a given M, this scheme maximizes the energy of the update in each iteration and thereby gives the closest possible performance to full filter update in terms of minimizing the mean squared error. Another tap selection scheme which exploits signal sparsity is the selective-partial-update (SPU) [19] tap selection scheme, where the N-tap adaptive filter is first divided into B blocks, which are then ranked according to the squared Euclidean norm of their respective tap-inputs. Based on this ranking, in each iteration the top $\lfloor B \cdot \frac{M}{N} \rfloor$ blocks, where $\lfloor \cdot \rfloor$ denotes the flooring operation, are selected to be updated. Many other schemes have been proposed which further improve performance by exploiting the sparseness of the echo path [38, 83]. Since sparseness of the echo path is more relevant for applications such as network echo cancellation [2], and not particularly relevant for the considered AEC application (as acoustic impulse responses are not particularly sparse), we will not consider such approaches in this paper.

Apart from large computational complexity, MAEC systems also suffer from other notable problems such as the misalignment problem [41–43]. Since in MAEC systems the different loudspeaker input signals are typically correlated with each other, the input covariance matrix may be ill-conditioned, possibly resulting in a large filter misalignment and a slow convergence speed. It should be realized that the misalignment problem is typically more severe in the context of speech communication systems, since the loudspeaker signals are obtained by filtering the same source (far-end speaker), as compared to surround-sound systems, where the loudspeaker signals may be independent of each other. The most common approach to tackle

the misalignment problem is to decorrelate the tap-inputs, for which several techniques have been proposed in literature [41, 42, 46]. Tap selection schemes such as the exclusive-maximum (XM) [47–49] have also been proposed to specifically tackle the misalignment problem for stereo AEC applications. The XM scheme improves the conditioning of the tap-input covariance matrix via exclusive updates of the two adaptive filters, i.e., in each iteration the same filter tap index is never selected in both channels. In this paper, however, we do not aim to solve the misalignment problem using tap selection schemes and do not claim to improve the misalignment performance for highly coherent loudspeaker signals, i.e., our main motivation is solely computational complexity reduction of MAEC systems.

As an alternative to time-domain adaptive filters, frequency-domain and subband adaptive filters are frequently used as they enable more efficient and frequency-dependent filter updates [1, 15, 51, 53, 56, 84]. Frequency-domain adaptive filtering algorithms, such as the fast least mean square (FLMS) [53], the partitioned block frequency-domain adaptive filtering (PB-FDAF) [84] and the multidelay block frequency-domain adaptive filtering (MDF) algorithm [56], are typically based on the overlap-save method [15, 51] and use the fast Fourier transform (FFT) to efficiently compute the required time-domain convolution and correlation operations. In [57], the M-Max tap selection scheme has been proposed for frequency-domain MDF algorithm. Alternatively, adaptive filtering can be performed using subband processing, where an analysis filterbank transforms the time-domain signals into the subband domain, the filter adaptation and processing is performed independently in each subband, and a synthesis filterbank is used to reconstruct the time-domain signals. In this paper, we will only consider subband adaptive filters. More specifically, we will use the well-known weighted overlap-add (WOLA) method [1, 61], i.e., using an FFT analysis filterbank to transform the (windowed) time-domain signals to the short-time Fourier transform (STFT) domain and an inverse FFT synthesis filterbank. Such a processing scheme provides a suitable compromise between computational complexity and latency, and enables to achieve a suitable time and frequency resolution.

In general, using a tap selection scheme may lead to a significant amount of processing overhead, primarily due to the required sorting effort. The computational savings obtained due to partial filter update are offset (and may even be exceeded in some cases) by the additional effort required for sorting. Compared to popular sorting algorithms such as the QUICKSORT routine [85], a more efficient fast running algorithm known as the SORTLINE routine [86] has been proposed for sorting vectors which contain many elements in common with a pre-sorted vector from a previous iteration, which is often the case with tap-input vectors from one iteration to the next.

In this paper, we propose and investigate different tap selection schemes in the subband domain for constrained partial updates of subband MAEC filters. Please note that in such a framework, the tap selection schemes operate on the magnitudes of the complex-valued STFT coefficients. Also, we consider the subband AEC filter in each channel to be composed of a number of sub-filters, i.e., one sub-filter per

subband. First, we extend the M-Max tap selection scheme proposed for complex-valued loudspeaker signals in [57] to the multichannel scenario, thereby applying the M-Max criterion across three dimensions, i.e., subbands, channels and filter length. Then, we present two new tap selection schemes which apply the M-Max criterion independently in each sub-filter across filter length only. The first scheme selects the same number of taps in each sub-filter, while the second scheme exploits the sparsity present in the loudspeaker signals across frequency and channels to select taps dynamically in the different sub-filters. Some preliminary results were obtained in [74] which indicated that signal sparsity present in real-world multichannel entertainment signals can be exploited to efficiently update the MAEC filters. The proposed tap selection schemes are then compared to the SPU tap selection scheme [19] in the subband domain[1].

The remainder of the paper is organized as follows. The signal model is presented in Section 2.3 and the different tap selection schemes considered are presented in Section 2.4. Section 2.5 presents a sparsity analysis for several synthetic and real-world multichannel signals, and the echo cancellation performance obtained when the different tap selection schemes are used. Section 2.6 discusses the computational effort required for the different tap selection schemes and the computational savings obtained when performing partial filter updates.

2.3 Signal model

We consider a loudspeaker–enclosure–microphone (LEM) system with R loudspeakers and a single microphone. The acoustic echo paths between the loudspeakers and the microphone are assumed to be time-invariant, such that the echo contribution from the r^{th} loudspeaker at discrete time index n is given by

$$d_r(n) = \sum_{v=0}^{V_r-1} h_r(v) \cdot x_r(n-v), \tag{2.1}$$

where x_r denotes the r^{th} input signal and h_r denotes the impulse response corresponding to the r^{th} acoustic echo path, with V_r denoting its length. Considering near-end speech signal s and near-end noise signal b, the microphone signal y is given as

$$y(n) = s(n) + d(n) + b(n), \tag{2.2}$$

where $d(n) = \sum_{r=1}^{R} d_r(n)$ denotes the total acoustic echo component.

For the subband-domain processing, an FFT analysis filterbank of order N_{FFT} is used to transform the (windowed) time-domain signals into the STFT domain, with

1 It should be noted that the XM tap selection scheme [47–49] cannot be straightforwardly implemented in the subband domain and extended to more than two channels.

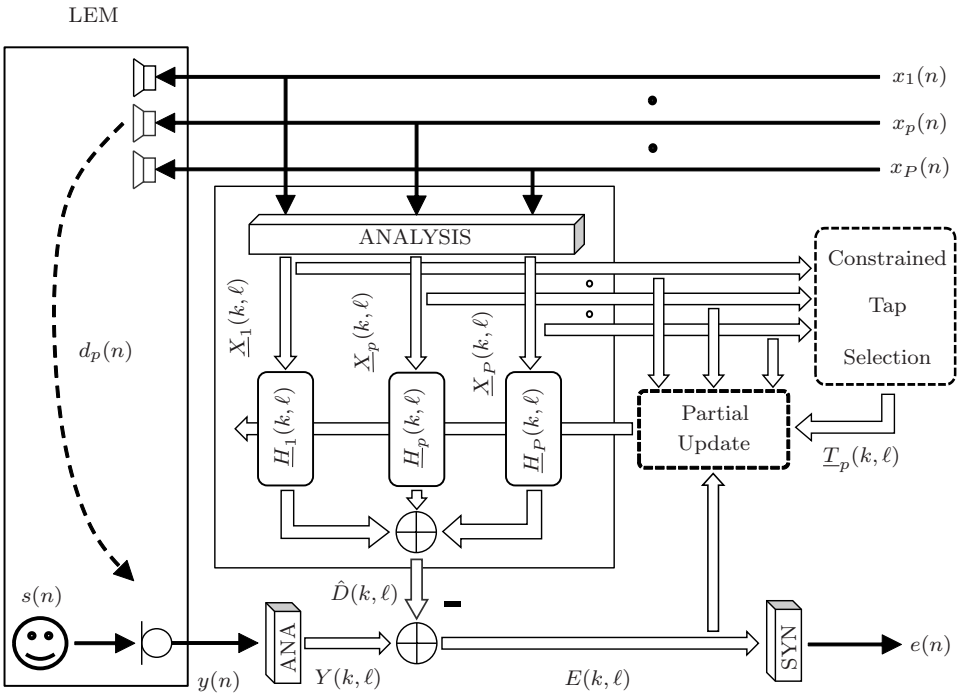

Fig. 2.1: Block diagram of the considered subband MAEC setup. Thin black arrows are used for signals processed in the time domain, while solid white arrows are used for signals processed in the subband domain.

the total number of subbands given by $K = \frac{N_{\mathrm{FFT}}}{2} + 1$. The STFT coefficient of the r-th input signal in the k-th subband and ℓ-th frame is computed as

$$X_r(k,\ell) = \sum_{m=0}^{N_{\mathrm{FFT}}-1} x_r(\ell \cdot F + m) \cdot W_{\mathrm{ana}}(m) \cdot e^{-j\frac{2\pi}{N_{\mathrm{FFT}}}km}, \tag{2.3}$$

where $j = \sqrt{-1}$, F denotes the frameshift and W_{ana} denotes the analysis window. In the remainder of the paper, the terms *reference channels* and *reference spectra* will be used to refer to the loudspeaker signals and their corresponding STFT coefficients, respectively.

The subband MAEC system is depicted in Figure 2.1 and consists of R adaptive filters, i.e., one corresponding to each reference channel, where each filter is composed of K sub-filters with L taps each. Thus, the total number of filter taps is given as

$$N = L \cdot K \cdot R, \tag{2.4}$$

i.e., L taps $\times$ K subbands $\times$ R channels.

The sub-filter for the k^{th} subband in the r^{th} channel is denoted as $\underline{\hat{H}}_r(k,\ell)$ and consists of L complex-valued coefficients

$$\underline{\hat{H}}_r(k,\ell) = \Big[\ \hat{H}_r^1(k,\ell) \quad \dots \quad \hat{H}_r^i(k,\ell) \quad \dots \quad \hat{H}_r^L(k,\ell)\ \Big]^T, \tag{2.5}$$

where $\hat{H}_r^i(k,\ell)$ denotes the i^{th} filter tap and $\cdot^T$ denotes the transpose operator. The tap-input vector to the sub-filter $\underline{\hat{H}}_r(k,\ell)$ also consists of L complex-valued spectral coefficients and is given as

$$\underline{X}_r(k,\ell) = \Big[\ X_r(k,\ell) \quad \dots \quad X_r(k,\ell-i+1) \quad \dots \quad X_r(k,\ell-L+1)\ \Big]^T. \tag{2.6}$$

The acoustic echo estimate for the r^{th} channel is generated by filtering the reference spectrum $\underline{X}_r(k,\ell)$ with the sub-filter $\underline{\hat{H}}_r(k,\ell)$

$$\hat{D}_r(k,\ell) = \underline{X}_r^H(k,\ell)\ \underline{\hat{H}}_r(k,\ell), \tag{2.7}$$

where $\cdot^H$ denotes the Hermitian operator. The total MAEC filter output is given as

$$\hat{D}(k,\ell) = \sum_{r=1}^{R} \hat{D}_r(k,\ell), \tag{2.8}$$

with the residual echo equal to

$$E(k,\ell) = Y(k,\ell) - \hat{D}(k,\ell), \tag{2.9}$$

where Y denotes the complex-valued spectrum of the microphone signal y, computed similarly to (2.3).

In order to reduce the computational complexity of the MAEC filter update in every frame, we will consider a partial update of $\underline{\hat{H}}_r(k,\ell)$ by updating only a subset $\mathcal{L}_r(k,\ell)$ of all L filter taps, where $\mathcal{L}_r(k,\ell)$ is an integer and is determined using a tap selection scheme (see Section 2.4). These tap selection schemes compute a vector

$$\underline{T}_r(k,\ell) = \Big[\ T_r^1(k,\ell) \quad \dots \quad T_r^i(k,\ell) \quad \dots \quad T_r^L(k,\ell)\ \Big]^T, \tag{2.10}$$

consisting of L binary-valued elements. If the element $T_r^i(k,\ell) = 1$, then the corresponding filter tap $\hat{H}_r^i(k,\ell)$ is selected to be updated, otherwise it is not. Thus, the sum of the elements of $\underline{T}_r(k,\ell)$ always satisfies

$$0 \le \sum_{i=1}^{L} T_r^i(k,\ell) = \mathcal{L}_r(k,\ell) \le L. \tag{2.11}$$

For updating $\underline{\hat{H}}_r(k,\ell)$, we use a variant of the normalized least mean squares (NLMS) algorithm [15], incorporating a partial filter update as shown below

$$\boxed{\underline{\hat{H}}_r(k,\ell+1) = \underline{\hat{H}}_r(k,\ell) + \left(\frac{\mu \cdot E^*(k,\ell)}{\mathcal{N}(k,\ell)+\epsilon}\right) \cdot \left\{\underline{T}_r(k,\ell) \odot \underline{X}_r(k,\ell)\right\},} \tag{2.12}$$

where μ denotes the (fixed) step-size, * denotes the complex-conjugate operator and $\odot$ denotes the element-wise multiplication operator. The step-size is normalized by the sum of the regularization parameter ϵ and the multichannel tap-input power

$$\mathcal{N}(k,\ell) = \sum_{r=1}^{R}\sum_{i=0}^{L-1} |X_r(k,\ell-i)|^2. \tag{2.13}$$

From hereon, we will refer to (2.12) as the partial update NLMS (PUNLMS) algorithm.

All tap selection schemes considered in this paper are based on the magnitudes of the tap-input vector $\underline{X}_r(k,\ell)$, i.e.,

$$\underline{\mathcal{X}}_r(k,\ell) = \Big[\ |X_r(k,\ell)| \ \ \ldots \ \ |X_r(k,\ell-i+1)| \ \ \ldots \ \ |X_r(k,\ell-L+1)| \ \Big]^T. \tag{2.14}$$

By stacking the vector $\underline{\mathcal{X}}_r(k,\ell)$ over all K subbands and R channels, we define the N-element vector

$$\underline{\mathbf{X}}(\ell) = \begin{bmatrix} \underline{\mathcal{X}}_1^T(1,\ell) & \ldots & \underline{\mathcal{X}}_1^T(K,\ell) & \ldots \\ \underline{\mathcal{X}}_r^T(1,\ell) & \ldots & \underline{\mathcal{X}}_r^T(K,\ell) & \ldots \\ \underline{\mathcal{X}}_R^T(1,\ell) & \ldots & \underline{\mathcal{X}}_R^T(K,\ell) & \end{bmatrix}^T, \tag{2.15}$$

containing the magnitudes of all MAEC filter tap-inputs. Similarly to (2.15), we define the N-element tap selection vector $\underline{\alpha}(\ell)$ by stacking the vector $\underline{T}_r(k,\ell)$ over all K subbands and R channels.

2.4 Tap selection schemes

In this section, we investigate and propose different tap selection schemes for designing the tap selection vector $\underline{\alpha}(\ell)$. All tap selection schemes exploit sparsity in $\underline{\mathbf{X}}(\ell)$ across one or more dimensions, i.e., frames, subbands and channels. A vector is considered *sparse* if a small number of its elements contain a large proportion of its energy. The terms *temporal*, *spectral* and *spatial* sparsity will be used to refer to sparsity present across frames, subbands and channels, respectively. For all consid-

ered schemes, we impose the constraint that in every frame exactly M taps across all $K \cdot R$ sub-filters are selected to be updated, with

$$M = \lfloor Q \cdot N \rfloor, \tag{2.16}$$

where $Q \in \mathcal{R}$ is a design parameter, with $0 \leq Q \leq 1$. Note that $Q = 0$ implies no filter update and $Q = 1$ implies full filter update. This also means that exactly M elements in the tap selection vector $\underline{\alpha}(\ell)$ are equal to 1, i.e.,

$$\boxed{\sum_{k=1}^{K} \sum_{r=1}^{R} \mathcal{L}_r(k, \ell) = M.} \tag{2.17}$$

The first tap selection scheme we investigate is the 3D M-Max scheme, which applies the M-Max criterion across the three dimensions of subbands, channels and filter length for selecting taps. Then we investigate the SPU scheme, which sorts the $K \cdot R$ sub-filters in each frame according to the squared Euclidean norm of their respective tap-inputs and then selects all L taps in the top $\lfloor \frac{M}{L} \rfloor$ sub-filters. Finally, we present two 1D M-Max schemes which apply the M-Max criterion only across the dimension of filter length, with the first scheme selecting the same number of taps in all sub-filters and the second scheme dynamically selecting taps in each sub-filter.

2.4.1 *3D M-Max (3DM) scheme*

The 3D M-Max tap selection scheme is an extension of the M-Max scheme proposed for the single-channel scenario in [57] to the multichannel scenario. Using this scheme, the filter taps corresponding to the M largest magnitude tap-inputs in every frame are selected to be updated by applying the M-Max criterion on the vector $\underline{\mathbf{X}}(\ell)$. The resulting tap selection vector $\underline{\alpha}(\ell)$ can then be unstacked to obtain the vectors $\underline{T}_r(k, \ell)$ corresponding to the $K \cdot R$ sub-filters. Implementing this scheme requires sorting the N-element vector $\underline{\mathbf{X}}(\ell)$ in every frame which is done efficiently using the QUICKSORT routine, requiring comparisons in the order of $\mathcal{O}(N \cdot \log_2 N)$ per frame.

As this scheme applies the M-Max criterion on the complete vector $\underline{\mathbf{X}}(\ell)$, it is able to exploit the *spectro-spatio-temporal* sparsity that may be present in the multichannel reference spectra, with the M selected taps distributed amongst the different sub-filters in every frame. For reference spectra with significant temporal, spatial and spectral diversity/non-stationarity, it is highly likely that each of the N filter taps are eventually updated at some stage. However, if the reference spectra exhibit stationarity and large spectral coloration and/or large inter-channel power difference, the M taps may be selected in only a small subset of the $K \cdot R$ sub-filters in every frame. This may result in the sub-filters in certain subbands and/or channels being completely ignored for a long time period, which may severely affect filter con-

vergence. This disadvantage of the 3DM scheme motivates us to look for schemes which do not completely ignore these sub-filters when allocating taps to be updated.

2.4.2 *SPU scheme*

In the SPU scheme [19], in each frame the $K \cdot R$ sub-filters are sorted according to the squared Euclidean norm of their respective tap-inputs

$$\eta_r(k,\ell) = ||\underline{\mathcal{X}}_r(k,\ell)||_2^2 = \sum_{i=0}^{L-1} |X_r(k,\ell-i)|^2. \tag{2.18}$$

All L taps in the top $\lfloor \frac{M}{L} \rfloor$ sub-filters are then selected to be updated, while no taps are selected in the remaining sub-filters. Hence, this scheme exploits the sparsity present in the multichannel reference spectra but suffers from the same problem as the 3DM scheme, i.e., it may completely ignore sub-filters in certain subbands and/or channels when the reference signals are spectrally coloured and stationary and/or exhibit large inter-channel power difference.

2.4.3 *1D M-Max schemes*

In this section, we present two tap selection schemes which apply the M-Max criterion only across the single dimension of filter length, thereby exploiting the temporal sparsity present in the multichannel reference spectra. Unlike the 3DM and SPU schemes, these two schemes are designed to not completely ignore the sub-filters with small magnitude tap-inputs when allocating taps to be updated. In both schemes, the M-Max criterion is applied on the L-element vector $\underline{\mathcal{X}}_r(k,\ell)$ for selecting taps in the sub-filter $\underline{\hat{H}}_r(k,\ell)$, with the number of taps selected given as

$$\boxed{\mathcal{L}_r(k,\ell) = \lfloor \psi_r(k,\ell) \cdot L \rfloor,} \tag{2.19}$$

where $\psi_r(k,\ell)$ is computed using two different criteria for the two schemes.

The fixed effort allocation (FEA) scheme selects the same number of filter taps in each sub-filter, thereby not exploiting spectral and spatial sparsity. On the other hand, the dynamic effort allocation (DEA) scheme selects filter taps in each sub-filter dynamically, aiming to exploit spectro-spatial sparsity while not ignoring sub-filters with small magnitude tap-inputs. It should be noted that $\psi_r(k,\ell)$ needs to satisfy the condition

$$0 \leq \psi_r(k,\ell) \leq 1, \tag{2.20}$$

as $\mathcal{L}_r(k,\ell)$ obviously cannot be larger than L. The vector $\underline{\mathcal{X}}_r(k,\ell)$ is sorted very efficiently using the SORTLINE routine, with the number of comparisons in the order of $\mathcal{O}(\log_2 L)$ per frame.

Substituting (2.16) and (2.19) into (2.17) gives

$$\sum_{k=1}^{K}\sum_{r=1}^{R}\lfloor \psi_r(k,\ell)\cdot L\rfloor = \lfloor Q\cdot N\rfloor. \tag{2.21}$$

Assuming no rounding errors when computing the flooring operation in (2.21), the constraint in (2.16) can be reformulated as

$$\boxed{\sum_{k=1}^{K}\sum_{r=1}^{R}\psi_r(k,\ell) = Q\cdot K\cdot R.} \tag{2.22}$$

2.4.3.1 *Fixed effort allocation (FEA)*

In the FEA scheme, the same number of filter taps are allocated to all $K\cdot R$ sub-filters in every frame, i.e.,

$$\boxed{\psi_r^F(k,\ell) = c,} \tag{2.23}$$

where the superscript ^{F} denotes the FEA scheme. Substituting (2.23) in (2.22) yields

$$c = Q. \tag{2.24}$$

Thus, in each sub-filter the filter coefficients corresponding to the $\lfloor Q\cdot L\rfloor$ largest magnitude tap-inputs are selected to be updated in every frame. Due to the same number of taps selected in all sub-filters, this scheme *does not* exploit the spectral and spatial sparsity present in the multichannel reference spectra.

2.4.3.2 *Dynamic effort allocation (DEA)*

In the DEA scheme, filter taps are dynamically allocated to the different sub-filters based on their respective tap-input content. We propose to allocate a larger number of taps in every frame to sub-filters with *relatively larger magnitude* tap-inputs, while not completely ignoring the sub-filters with smaller magnitude tap-inputs. Thus, the DEA scheme aims to combine the advantages of the 3DM and the FEA schemes while avoiding their disadvantages, i.e., exploiting the *spectro-spatial sparsity* present in the multichannel reference spectra, while not ignoring the sub-filters with small magnitude tap-inputs.

In general, in the DEA scheme the number of filter taps allocated to the sub-filter for the k^{th} subband in the r^{th} channel is based on the corresponding tap-input content, which can be quantified by

$$\phi_r(k,\ell) = ||\underline{\mathcal{X}}_r(k,\ell)||_p^p = \sum_{i=0}^{L-1} |X_r(k,\ell-i)|^p, \tag{2.25}$$

where $||\cdot||_p$ denotes the l_p-norm for $p > 0$. Hence, sub-filters with larger magnitude tap-inputs will have larger values of $\phi_r(k,\ell)$ as compared to sub-filters with smaller magnitude tap-inputs. Note that for simplicity, we have used $p = 1$. The factor $\psi_r(k,\ell)$ in (2.19) is then computed as

$$\psi_r^G(k,\ell) = \min\left\{f\Big(\phi_r(k,\ell)\Big), 1\right\}, \tag{2.26}$$

where the superscript ^{G} denotes the generic form of the DEA scheme, the function $f(\cdot)$ depends on the used tap allocation criterion and the minimum operator is required to satisfy the condition in (2.20). The number of taps selected in the sub-filter $\underline{\hat{H}}_r(k,\ell)$ is finally determined by substituting (2.26) in (2.19).

We propose to design the function $f(\cdot)$ based on the simple criterion that sub-filters with $\phi_r(k,\ell)$ above a certain threshold $\phi_{\text{th}}(\ell)$ get L filter taps selected, while all other sub-filters get a number proportional to $\phi_r(k,\ell)$, i.e.,

$$f\Big(\phi_r(k,\ell)\Big) = \frac{\phi_r(k,\ell)}{\phi_{\text{th}}(\ell)}. \tag{2.27}$$

Choosing an appropriate value for the threshold $\phi_{\text{th}}(\ell)$ is quite important. On the one hand, choosing a low value could result in a large number of sub-filters having L taps updated, which potentially dilutes the extent to which spectro-spatial sparsity is exploited for tap allocation. On the other hand, choosing a large value could result in a large number of sub-filters being completely ignored. Hence, we propose to use the average value of $\phi_r(k,\ell)$ across all subbands and channels, i.e.,

$$\phi_{\text{th}}(\ell) = \phi_{\text{avg}}(\ell) = \frac{1}{K \cdot R} \sum_{k=1}^{K} \sum_{r=1}^{R} \phi_r(k,\ell). \tag{2.28}$$

However, when using the function in (2.27) with the threshold in (2.28), it cannot be guaranteed that the constraint in (2.22) is satisfied in every frame. Since $\min(a,1) \leq a$ for any real number $a \in \mathcal{R}$, it can be easily shown that

$$\begin{aligned} \sum_{k=1}^{K} \sum_{r=1}^{R} \psi_r^G(k,\ell) &\leq \sum_{k=1}^{K} \sum_{r=1}^{R} f\Big(\phi_r(k,\ell)\Big) \\ &\leq \frac{1}{\phi_{\text{avg}}(\ell)} \cdot \sum_{k=1}^{K} \sum_{r=1}^{R} \phi_r(k,\ell), \end{aligned} \tag{2.29}$$

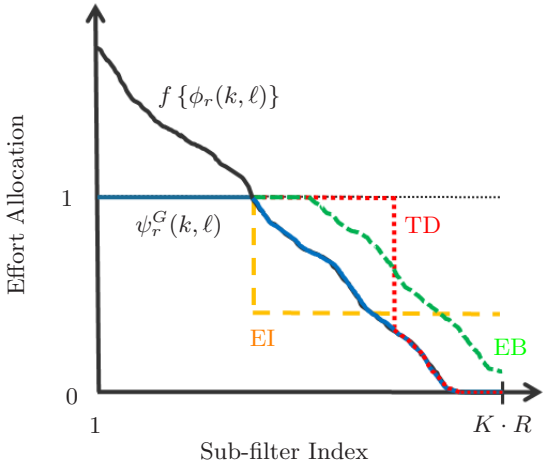

Fig. 2.2: Exemplary function $f\Big(\phi_r(k,\ell)\Big)$ and corresponding $\psi_r^G(k,\ell)$, plotted in sorted order of highest to lowest values, along with different criteria for modifying $\psi_r^G(k,\ell)$ in case $M_{\mathrm{G}}(\ell) < Q \cdot K \cdot R$.

such that

$$M_{\mathrm{G}}(\ell) = \sum_{k=1}^{K} \sum_{r=1}^{R} \psi_r^G(k,\ell) \leq K \cdot R. \tag{2.30}$$

Thus, it is not guaranteed that $M_{\mathrm{G}}(\ell)$ is equal to $Q \cdot K \cdot R$, and hence the constraint in (2.22) may not always be satisfied.

We will now distinguish 2 cases, i.e., $M_{\mathrm{G}}(\ell) < Q \cdot K \cdot R$ and $M_{\mathrm{G}}(\ell) > Q \cdot K \cdot R$, and discuss how to adjust the filter tap allocation in order to satisfy the constraint.

- Case 1: $M_{\mathrm{G}}(\ell) < Q \cdot K \cdot R$

  Figure 2.2 shows an exemplary function $f\Big(\phi_r(k,\ell)\Big)$ (black curve) and corresponding $\psi_r^G(k,\ell)$ (blue curve) plotted for all $K \cdot R$ sub-filters for the case $M_{\mathrm{G}}(\ell) < Q \cdot K \cdot R$, sorted from largest to smallest value in terms of $\phi_r(k,\ell)$. Please note that the area under the black curve is equal to $K \cdot R$, while the area under the blue curve is equal to $M_{\mathrm{G}}(\ell)$. In order to satisfy the constraint in (2.22), the surplus effort $Q \cdot K \cdot R - M_{\mathrm{G}}(\ell)$ needs to be redistributed amongst the sub-filters for which $\psi_r^G(k,\ell) < 1$. In order to do so, different criteria can be used for modifying $\psi_r^G(k,\ell)$:

  - *Trickle Down* (TD): When using this criterion (red), the surplus effort is redistributed via the trickle-down procedure, i.e., the sub-filters are filled up in sorted order of $\psi_r^G(k,\ell)$. Allocating taps in this way respects the spectro-spatial sparsity present in the tap-inputs, but would most likely *completely ignore* sub-filters with the smallest magnitude tap-inputs.

- *Equal Income* (EI): When using this criterion (orange), the same number of taps are allocated in all sub-filters for which $\psi_r^G(k,\ell) < 1$. This has the beneficial effect that no sub-filters are ignored, but has the detrimental effect that the spectro-spatial sparsity present in the tap-inputs would most likely *not* be exploited for tap allocation.
- *Equal Bonus* (EB): When using this criterion (green), the surplus effort is redistributed equally amongst all sub-filters for which $\psi_r^G(k,\ell) < 1$. Allocating taps in this way respects the spectro-spatial sparsity present in the tap-inputs while making sure that all sub-filters get a few taps updated.

Since the EB criterion attains a balance between exploiting spectro-spatial sparsity and not completely ignoring sub-filters, we decide to use this criteria in our proposed DEA scheme when $M_{\mathrm{G}}(\ell) < Q \cdot K \cdot R$, i.e.,

$$\psi_r^D(k,\ell) = \{1 - \gamma(\ell)\} + \gamma(\ell) \cdot \psi_r^G(k,\ell), \tag{2.31}$$

where the superscript ^{D} denotes the proposed DEA scheme. The constant $\gamma(\ell)$ can be computed by substituting (2.31) into (2.22), yielding

$$\gamma(\ell) = \frac{K \cdot R - Q \cdot K \cdot R}{K \cdot R - M_{\mathrm{G}}(\ell)}. \tag{2.32}$$

Thus, each sub-filter has a minimum of $\lfloor \{1 - \gamma(\ell)\} \cdot L \rfloor$ taps selected in the ℓ^{th} frame.

- Case 2: $M_{\mathrm{G}}(\ell) > Q \cdot K \cdot R$

Similarly to Figure 2.2, Figure 2.3 shows an exemplary function $f\Big(\phi_r(k,\ell)\Big)$ (black curve) and corresponding $\psi_r^G(k,\ell)$ (blue curve) for the case $M_{\mathrm{G}}(\ell) > Q \cdot K \cdot R$. In order to satisfy the constraint, different criteria can be used for modifying $\psi_r^G(k,\ell)$:

- *Tax the Poor* (TP): When using this criterion (red), the constraint is satisfied by decreasing the number of taps allocated to sub-filters with the lowest $\psi_r^G(k,\ell)$. Such a scheme typically results in highly unequal tap allocation, with all taps reserved for a small number of sub-filters with the largest magnitude tap-inputs.
- *Tax the Rich* (TR): When using this criterion (orange), the constraint is satisfied by decreasing the number of taps allocated to sub-filters with the highest $\psi_r^G(k,\ell)$. This scheme has the beneficial effect that the majority of sub-filters are not ignored when allocating taps but has the detrimental effect that the spectro-spatial sparsity present in the tap-inputs is most likely *not* exploited for tap allocation.

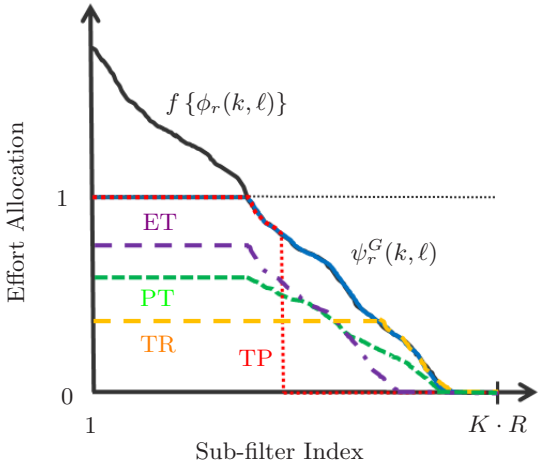

Fig. 2.3: Exemplary function $f\Big(\phi_r(k,\ell)\Big)$ and corresponding $\psi_r^G(k,\ell)$, plotted in sorted order of highest to lowest values, along with different criteria for modifying $\psi_r^G(k,\ell)$ in case $M_{\mathrm{G}}(\ell) > Q \cdot K \cdot R$.

- *Equal Tax* (ET): When using this criterion (violet), the constraint is satisfied by decreasing the same number of taps from all $K \cdot R$ sub-filters. At first, this looks like a fair way of subtracting taps as it respects the spectro-spatial sparsity in the tap-inputs. However, it can be observed that this criterion ignores sub-filters with the smallest magnitude tap-inputs, as it takes away any small number of taps that may have been previously allocated to them.
- *Proportionate Tax* (PT): When using this criterion (green curve), the constraint is satisfied by uniformly scaling down the number of allocated taps in the different sub-filters. Allocating taps in this way respects the spectro-spatial sparsity present in the tap-inputs, while ensuring that lesser number of taps are reduced from sub-filters with smaller $\psi_r^G(k,\ell)$.

Since the PT criterion attains a good balance between exploiting spectro-spatial sparsity and not completely ignoring sub-filters, we decide to use this criterion in our proposed DEA scheme when $M_{\mathrm{G}}(\ell) > Q \cdot K \cdot R$, i.e.,

$$\psi_r^D(k,\ell) = \delta(\ell) \cdot \psi_r^G(k,\ell), \tag{2.33}$$

where the constant $\delta(\ell)$ can be computed by substituting (2.33) into (2.22), yielding

$$\delta(\ell) = \frac{Q \cdot K \cdot R}{M_{\mathrm{G}}(\ell)}. \tag{2.34}$$

The proposed DEA scheme can thus be summarized as

$$\psi_r^D(k,\ell) = \begin{cases} \{1-\gamma(\ell)\} + \gamma(\ell)\cdot\psi_r^G(k,\ell), & \text{if } M_{\mathrm{G}}(\ell) < Q\cdot K\cdot R \\ \\ \delta(\ell)\cdot\psi_r^G(k,\ell), & \text{if } M_{\mathrm{G}}(\ell) \geq Q\cdot K\cdot R. \end{cases} \tag{2.35}$$

The number of taps selected to be updated in the sub-filter $\underline{\hat{H}}_r(k,\ell)$ using the DEA scheme is finally determined by substituting (2.35) into (2.19).

2.5 Simulations, results and discussion

In this section, we present the reference signals and algorithmic parameters used, as well as the different metrics used to analyze signal sparsity, tap selection and echo cancellation performance. We perform a sparsity analysis of the multichannel reference signals, individually across the three dimensions of subbands, channels and filter length, as well as jointly across multiple dimensions. We then analyze the effect of using the different tap selection schemes on the echo cancellation performance obtained for the different types of reference signals used.

2.5.1 *Signals and algorithmic parameters*

In our simulations, we use time-domain reference signals at a sampling frequency of $f_s = 16$ kHz. The different reference signals used can be divided into two categories:

- Synthetic signals
  - Mono brown and white noise signals, i.e., signals whose power densities change at the rate of -6 and 0 dB/octave, respectively
  - Stereo white noise signal
- Real-world signals
  - Mono speech signals (TIMIT database)
  - Surround-sound movie signals (Dolby Digital 5.0 format)
  - Surround-sound concert signals (Dolby Digital 5.0 format)

The acoustic impulse responses have been measured in a room with $T_{60} \approx 550$ ms, with the microphone and the 5 loudspeakers placed on a circle of 3m radius. The microphone was placed at a height of 1.2m, the centre (C) loudspeaker was placed directly 0.85m below the microphone, the front left (FL) and right (FR) loudspeakers were placed at the same height and 30° either side of the microphone, and the side left (SL) and right (SR) loudspeakers were placed 0.4m above and 110° either side of the microphone, respectively. The acoustic echo signal d_r is obtained

by convolving the reference signal x_r with the corresponding impulse response h_r for $V_r = 200$ ms. We assume no near-end speech signal ($s(n) = 0$) and no additive near-end noise signal ($b(n) = 0$) for our simulations. For the mono reference signals, we use the impulse response corresponding to the C loudspeaker only, while for the stereo white noise signal, we use the impulse responses corresponding to the FL and FR channels. The time-domain signals have been transformed into the subband domain using STFT processing with $N_{\text{FFT}} = 512$ (i.e., $K = 257$) using a Hanning window and an overlap of 75%. We use a filter length $L = 20$ for the MAEC filters, which corresponds to $N_{\text{FFT}} \cdot \{1 + 0.25 \cdot (L-1)\}$ samples or 184 ms. For updating the MAEC filters, a fixed step-size of $\mu = 0.1$ and regularization parameter of $\epsilon = 10^{-60}$ have been used.

2.5.2 *Performance measures*

Here, we present the different metrics used to analyze the sparsity present in the reference spectra, to analyze the performance of the different tap selection schemes in exploiting signal sparsity and to measure echo cancellation performance.

2.5.2.1 *Sparsity metric*

To analyze the sparsity in the multichannel reference spectra across subbands, channels and frames, different metrics exist, such as the l_0-norm, the l_1-norm, the Gini index [87] and the Hoyer metric [88]. For an N-element (non-zero) vector $\underline{u} = [u_0 \dots u_{N-1}]$, where the elements are sorted in order of magnitude $|u_0| \leq \dots \leq |u_{N-1}|$, the Gini index is defined as

$$g(\underline{u}) = 1 - 2 \cdot \sum_{j=0}^{N-1} \left(\frac{N - j - 0.5}{N} \right) \cdot \frac{|u_j|}{\sum_{i=0}^{N-1} |u_i|}. \tag{2.36}$$

On the one hand, for the extreme case where $|u_0| = \dots = |u_{N-1}|$, i.e., no sparsity in $\underline{u}$, $g(\underline{u}) = 0$. On the other hand, for the extreme case where $|u_0| = \dots = |u_{N-2}| = 0$ and $|u_{N-1}| \neq 0$, i.e., very high sparsity in $\underline{u}$, $g(\underline{u}) = 1 - \frac{1}{N}$, which for a large value of N is approximately equal to 1. Thus, the sparser the vector, the higher the Gini index.

Furthermore, the Gini index exhibits the following properties:

- Limited range: $0 \leq g(\underline{u}) \leq 1$
- Scaling invariance: $g(a \cdot \underline{u}) = g(\underline{u})$, $\forall\ a \in \mathcal{R}$
- Sensitivity to addition: $g(a + \underline{u}) < g(\underline{u})$, $\forall\ a \in \mathcal{R}, a > 0$
- Cloning invariance: $g(\underline{u}) = g([\underline{u}\ \underline{u}]) = g([\underline{u}\ \underline{u}\ \underline{u}])$
- Sensitivity to zero-padding: $g([\underline{u}\ 0]) > g(\underline{u})$

The cloning invariance property allows a fair comparison of the sparsity of vectors with different number of elements. This is an important consideration, as we want to compare the sparsity of the reference spectra across the different dimensions of subbands, channels and frames. Note that the oft-used Hoyer metric does not exhibit this invariance and is hence not suited for comparing vectors with different number of elements.

2.5.2.2 *Tap selection performance*

In order to quantify the *closeness* of a tap selection scheme to full tap selection, we use the so-called Closeness Measure [48, 49] which is defined as the ratio of the energy of the M selected tap-inputs to the energy of all tap-inputs, i.e.,

$$\xi\Big(\underline{\alpha}(\ell), \underline{\mathbf{X}}(\ell)\Big) = \frac{||\underline{\alpha}(\ell) \odot \underline{\mathbf{X}}(\ell)||_2^2}{||\underline{\mathbf{X}}(\ell)||_2^2}. \tag{2.37}$$

For full filter update, i.e., $\underline{\alpha}(\ell) = \underline{1}$, we obviously obtain $\xi = 1$. For a given Q, the 3DM scheme maximizes the Closeness Measure in every frame, as it selects the M largest magnitude tap-inputs. The expectation and assumption is that the tap selection scheme yielding the largest Closeness Measure also results in the smallest difference in AEC performance compared to updating the filters using full tap selection.

2.5.2.3 *Echo cancellation performance*

The echo cancellation performance is evaluated using the echo return loss enhancement (ERLE) [1], which is defined as

$$\mathrm{ERLE}(n) = 10 \cdot \log_{10} \frac{E\left[d^2(n)\right]}{E\left[\Big(d(n) - \hat{d}(n)\Big)^2\right]}, \tag{2.38}$$

where $\hat{d}(n)$ is the time-domain signal corresponding to the total MAEC filter output $\hat{D}(k, \ell)$ and $E[\cdot]$ denotes the statistical expectation operator. In practice, the ERLE is computed by approximating the expectation operator with the current sample value. The speed of convergence of the MAEC filters is assessed using the t_{20} metric, which is the time required for the ERLE to reach 20 dB.

2.5.3 *Sparsity analysis*

In this section, we present an example to illustrate the amount of sparsity typically present in real-world multichannel spectra across subbands, channels and frames, and also jointly across multiple dimensions. Figure 2.4 (a) depicts the waveform of a 10s segment from the soundtrack of a 5-channel movie signal, with the spectrograms

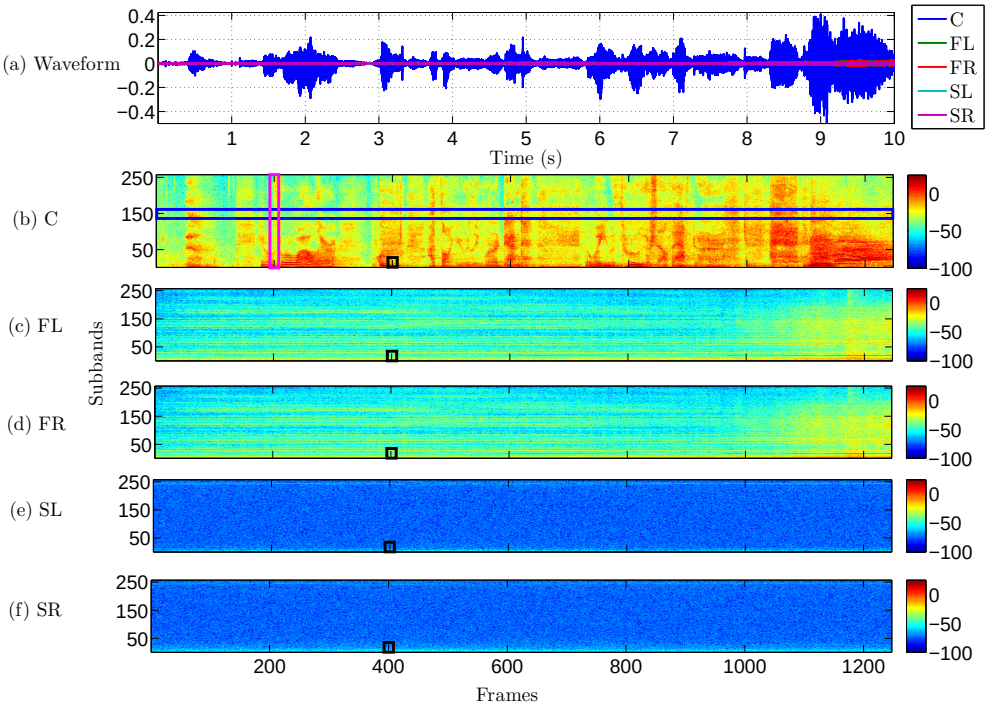

Fig. 2.4: (a) Waveform of a 10s segment from the soundtrack of a 5-channel movie signal, with different channels distinguished by color; magnitude spectrogram of (b) centre (C), (c) front left (FL), (d) front right (FR), (e) side left (SL) and (f) side right (SR) channels, respectively.

of the C, FL, FR, SL and SR channels shown in the subplots below. Each magnitude spectrogram is composed of $K = 257$ subbands and $T = 1247$ frames. In this movie signal, the centre channel contains the speech content, while the surround-sound channels contain the background score.

From these spectrograms, we first analyze the sparsity across subbands (spectral sparsity), across frames (temporal sparsity) and across channels (spatial sparsity). The Gini index for spectral sparsity in each channel is computed in every frame on a vector of K spectral coefficients, as exemplarily shown in Figure 2.4 (b) for the centre channel using the magenta box in frame 200. Similarly, the Gini index for temporal sparsity in each channel is computed on a vector of T spectral coefficients in every subband, as shown using the blue box for subband 150. The Gini index for spatial sparsity in each subband and frame is computed on a vector of R spectral coefficients, as exemplarily shown using the black boxes for the first subband in frame 400. The Gini indices so obtained for spectral, temporal and spatial sparsity are shown in Figure 2.5 (a), (b) and (c), respectively. It can be observed that the multichannel reference spectra displays a fairly high amount of sparsity across all the three dimensions individually, with Gini indices on average above 0.5 (except

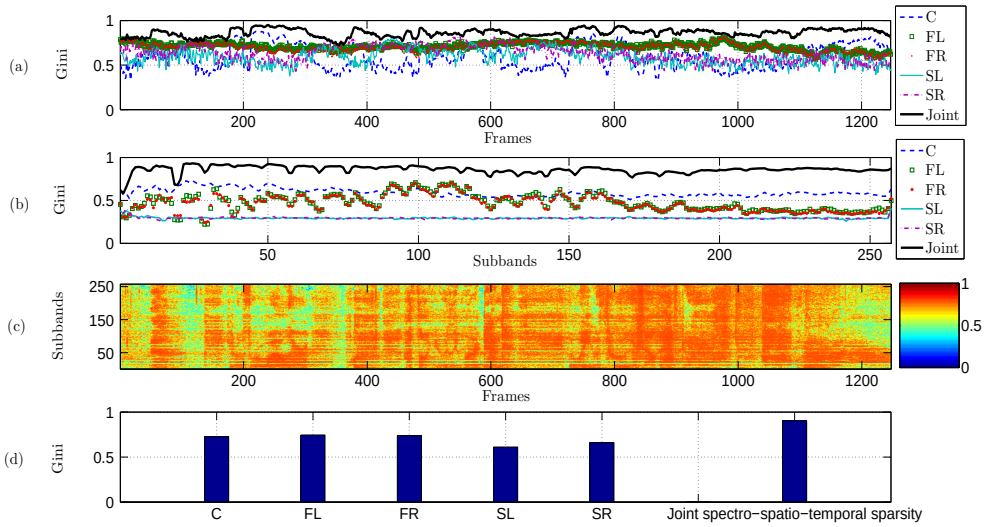

Fig. 2.5: Gini indices for a 10s segment from the soundtrack of a 5-channel movie signal; (a) spectral sparsity in each channel and joint spectro-spatial sparsity, (b) temporal sparsity in each channel and joint spatio-temporal sparsity, (c) spatial sparsity in each subband and frame, (d) joint spectro-temporal sparsity in each channel and joint spectro-spatio-temporal sparsity.

for temporal sparsity in the surround-sound channels). The centre channel displays higher temporal sparsity as compared to the surround-sound channels as it contains time-varying speech content, while the surround-sound channels contain the background score, which varies slowly with time.

Additionally, we analyze the sparsity present in the spectra *jointly* across multiple dimensions. In Figure 2.5 (a), the black curve displays the Gini index for the joint spectro-spatial sparsity, computed in every frame on a vector with $K \cdot R$ spectral coefficients. Similarly, in Figure 2.5 (b), the black curve displays the Gini index for the joint spatio-temporal sparsity, computed in every subband on a vector with $R \cdot T$ spectral coefficients. The Gini index for the joint spectro-temporal sparsity in each channel is computed by processing the magnitude spectrogram of that channel and is plotted in Figure 2.5 (d), along with the joint spectro-spatio-temporal sparsity for all $K \cdot R \cdot T$ coefficients. From this figure, it can be clearly observed that the multichannel reference spectra exhibit even higher levels of sparsity when analyzed across multiple dimensions, with Gini indices on average above 0.85. This provides the motivation to exploit sparsity *jointly* across subbands, channels and frames for the purpose of tap selection.

Figure 2.6 shows the Gini indices for the joint spectro-spatio-temporal sparsity for the different considered reference signals. The stereo white noise signal is chosen to be spatially sparse, with an inter-channel broadband power ratio of 20 dB. Firstly, it can be observed for the synthetic signals that the spectrally colored brown noise

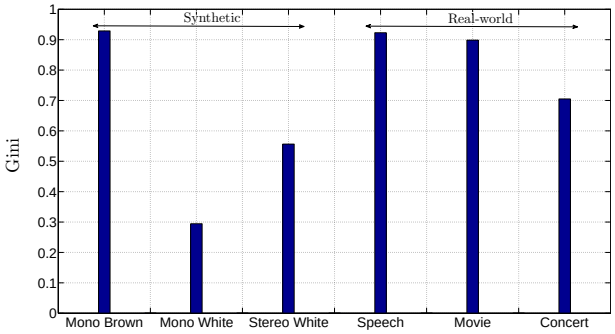

Fig. 2.6: Gini indices for joint spectro-spatio-temporal sparsity for different reference signals.

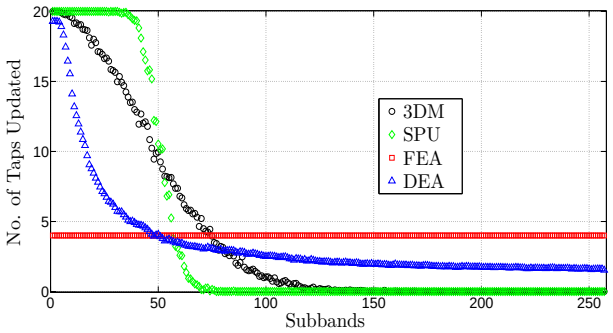

Fig. 2.7: Number of taps selected in each subband when using the 3DM, SPU, FEA and DEA tap selection schemes for a mono brown noise signal ($Q = 0.2$).

signal and the stereo white noise signal are obviously more sparse than the mono white noise signal. Secondly, it can be observed that typical real-world signals such as mono speech and 5-channel movie and concert signals also display high amounts of sparsity.

2.5.4 *Analysis of tap selection schemes for synthetic signals*

In this section, we analyze the effect of using the constrained tap selection schemes from Section 2.4 (3DM, SPU, FEA and DEA) for synthetic signals.

2.5.4.1 *Effect of spectral coloration*

For the different tap selection schemes, Figure 2.7 shows the number of taps selected in each subband when using a mono brown signal with $Q = 0.2$. For the 3DM and SPU schemes, a larger number of taps are selected in the low-frequency subbands which contain the larger magnitude tap-inputs, while the high-frequency

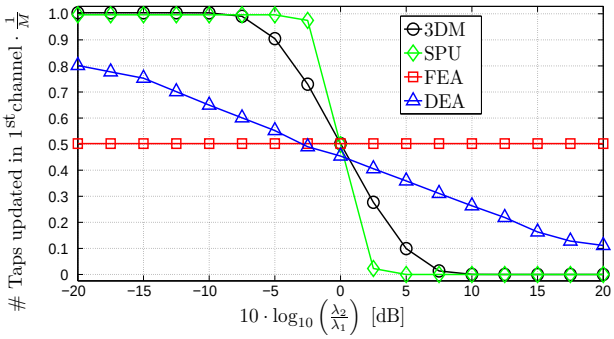

Fig. 2.8: Effect of the inter-channel power ratio of a stereo white noise signal on the number of taps allocated to the sub-filters in the first channel (as a fraction of M taps allocated to both channels) when using the 3DM, SPU, FEA and DEA tap selection schemes ($Q = 0.2$).

subbands with the smallest magnitude tap-inputs get no taps selected. Since the FEA scheme does not exploit spectral sparsity, it allocates an equal number of taps in all sub-filters irrespective of the signal content. The proposed DEA scheme achieves a balance by allocating more taps to sub-filters with larger magnitude tap-inputs (thereby exploiting spectral sparsity), while not completely ignoring the sub-filters with the smallest magnitude tap-inputs.

2.5.4.2 *Effect of inter-channel power ratio*

We now consider a stereo white noise signal, where the broadband power of the first and the second channel is denoted as λ_1 and λ_2, respectively. Figure 2.8 shows the effect of the inter-channel power ratio $\frac{\lambda_2}{\lambda_1}$ on the number of taps selected in the sub-filters of the first channel (as a fraction of the M taps selected in both channels) for the different tap selection schemes with $Q = 0.2$. When using the 3DM and SPU schemes, for $\lambda_1 > \lambda_2$, the sub-filters in the first channel get the majority of the M taps selected. Thus, both schemes are highly spatially selective, as hardly any taps of the sub-filters in the less dominant reference channel are updated (e.g., for the SPU scheme when the inter-channel power ratio is larger than 5 dB and for the 3DM scheme when the inter-channel power difference ratio is larger than 10 dB). Since the FEA scheme does not exploit spatial sparsity, it allocates an equal number of taps to the sub-filters in the first and the second channel (i.e., $\frac{M}{2}$ taps each), irrespective of the inter-channel power ratio. The proposed DEA scheme achieves a balance by allocating more taps to the sub-filters in the dominant reference channel (thereby exploiting spatial sparsity), while not completely ignoring the channel with the smaller magnitude tap-inputs.

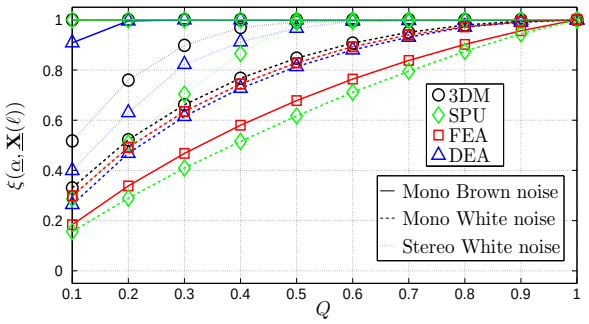

Fig. 2.9: Closeness Measure as a function of Q for mono brown, mono white and stereo white noise signals for different tap selection schemes.

2.5.4.3 *Closeness measure*

For different values of Q, Figure 2.9 depicts the Closeness Measure ξ obtained when using the different tap selection schemes for mono brown, mono white and stereo white noise signals. For the stereo white noise signal, an inter-channel power ratio of 20 dB has been chosen. This figure shows how close the different tap selection schemes are to full tap selection in terms of the energy of the selected tap-inputs. By design, the 3DM scheme maximizes the Closeness Measure for a given Q, and hence yields the highest values for each signal. For a highly sparse signal such as the mono brown signal, a very high value for the Closeness Measure (≈ 1) is obtained for the 3DM scheme even when only 10% of the total filter taps are selected (i.e., $Q = 0.1$). This means that just 10% of the tap-inputs contain almost the entire energy. For the least sparse mono white noise signal, low values of the Closeness Measure are obtained for all schemes, especially for the SPU scheme. For example, for $Q = 0.5$, a Closeness Measure of about 0.85 is obtained for the 3DM, FEA and DEA schemes, whereas a Closeness Measure of about 0.6 is obtained for the SPU scheme. The Closeness Measure values obtained for the stereo white signal for all schemes lie in between those obtained for the more sparse mono brown noise signal and the less sparse mono white noise signal, except for the FEA scheme, which yields the same values as for the mono white noise signal. The SPU scheme gives high values for highly sparse signals and very low values for signals with low amounts of sparsity, while the proposed DEA scheme performs similarly to the 3DM scheme for highly sparse signals and similarly to the FEA scheme for signals with low amounts of sparsity.

2.5.4.4 *ERLE and t_{20}*

As shown by the previous experiments, depending on the spectral coloration and the inter-channel power ratio of the reference signals, each considered tap selection scheme results in a different distribution of the selected taps across subbands and channels, and a different Closeness Measure. Hence, it is to be expected that the

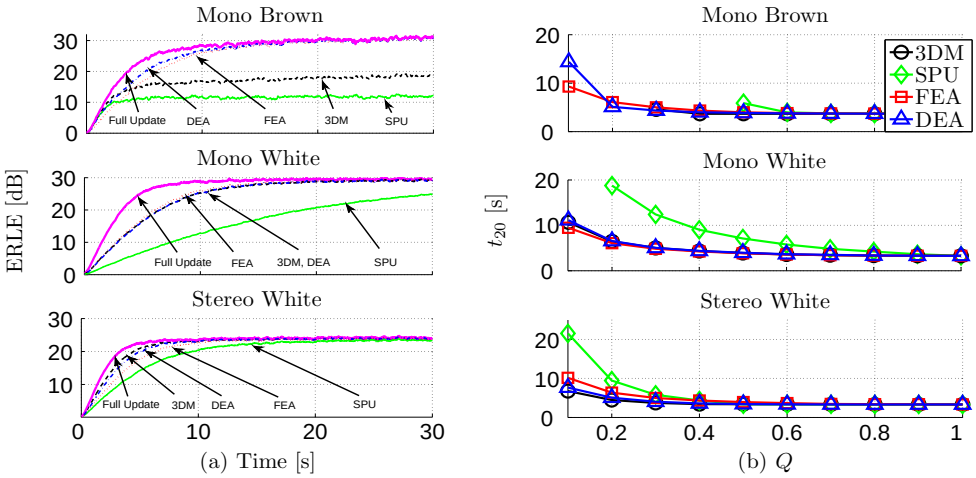

Fig. 2.10: (a) ERLE convergence curves for full filter update ($Q = 1$) and for different tap selection schemes ($Q = 0.2$), and (b) t_{20} values for different values of Q (for mono brown, mono white and stereo white noise signals).

tap selection schemes have an influence on the overall acoustic echo cancellation performance, i.e., ERLE and speed of filter convergence.

For mono brown, mono white and stereo white noise (inter-channel power ratio of 20 dB) signals, Figure 2.10 (a) shows the ERLE convergence curves for the 3DM, SPU, FEA and DEA tap selection schemes ($Q = 0.2$), compared to full filter update ($Q = 1$). Figure 2.10 (b) shows the corresponding t_{20} values for different values of the parameter Q. It can be observed that for signals with a high amount of spectral sparsity, such as the mono brown noise signal, the DEA scheme yields the best echo cancellation performance, while the 3DM and SPU schemes yield the poorest performance despite obtaining the highest values for the Closeness Measure. This is due to the highly spectrally selective nature of the 3DM and SPU schemes (discussed in Section 2.5.4.1), i.e., the sub-filters with the smallest magnitude tap-inputs do not have taps updated in every frame, resulting in very slow convergence of these sub-filters and thus negatively affecting the overall echo cancellation performance. For the least sparse mono white noise signal, it can be observed that the 3DM, FEA and DEA schemes yield similar echo cancellation performance, while the SPU again yields the poorest performance. This may be due to the fact that the SPU scheme is the only one which completely ignores entire subbands when updating the filters, while the other schemes may allocate a few taps to each subband when the reference signal has a low amount of sparsity. For the spatially sparse stereo white noise signal, the DEA scheme performs better than the FEA scheme, both in terms of the converged ERLE value as well as the t_{20} values. For all considered signals, the ERLE and t_{20} values obtained by the proposed DEA scheme for $Q = 0.2$ are very similar to those obtained for full filter update. Thus, the DEA scheme gives very similar echo cancellation performance to full filter update even when only 20% of the total MAEC filter taps are updated in every frame.

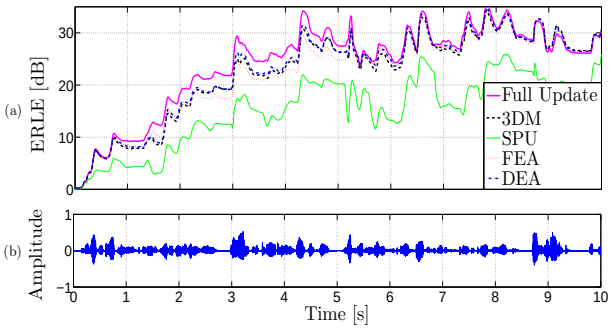

Fig. 2.11: (a) ERLE curves obtained for full update ($Q = 1$) and for different tap selection schemes ($Q = 0.2$) for a 10s segment of a mono speech signal; (b) waveform of the 10s segment of a mono speech signal.

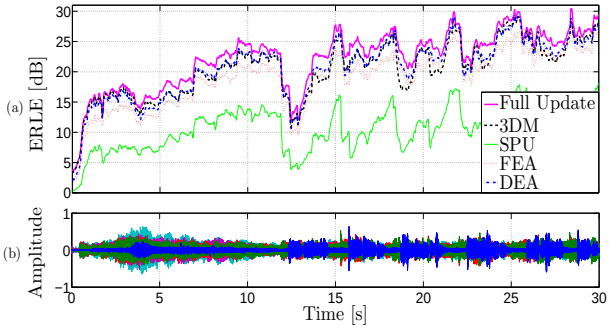

Fig. 2.12: (a) ERLE curves obtained for full update ($Q = 1$) and for different tap selection schemes ($Q = 0.2$) for a 30s segment of a 5-channel concert signal; (b) waveform of the 30s segment of a 5-channel concert signal.

2.5.5 *Analysis of tap selection schemes for real-world signals*

Contrary to the synthetic (stationary) signals in the previous section, in this section we investigate the effect of using constrained tap selection schemes on the echo cancellation performance for (non-stationary) real-world signals.

For a mono speech signal, Figure 2.11 shows the ERLE curves obtained when the MAEC filters are updated using the different tap selection schemes for $Q = 0.2$ and for full filter update ($Q = 1$) for a period of 10s. For this signal, we find that even when only 20% of all filter taps are updated in every frame, both the 3DM scheme and the proposed DEA scheme typically perform as well as full filter update in terms of ERLE, with the FEA scheme performing slightly worse (about 1-2 dB). On the other hand, the SPU scheme performs significantly worse, yielding about 7-8 dB deterioration in terms of ERLE.

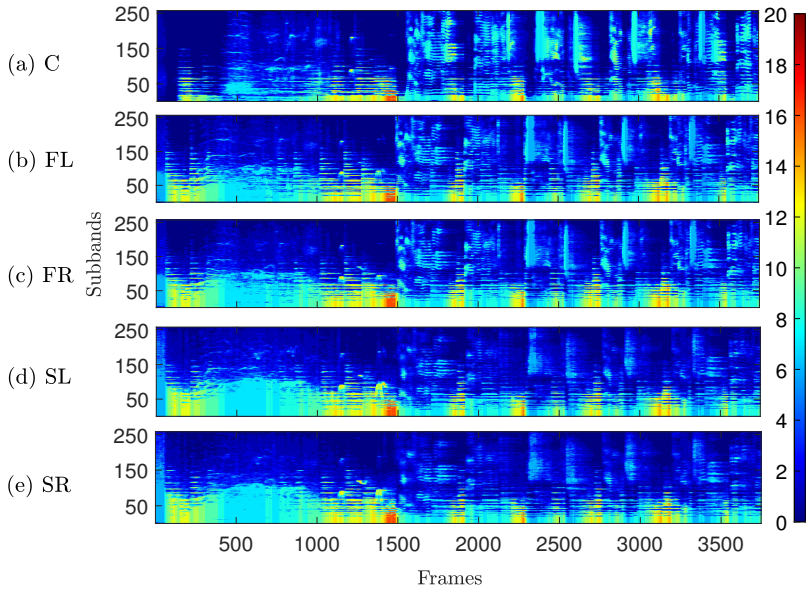

Fig. 2.13: Number of taps $\mathcal{L}_r(k,\ell)$ updated in the different sub-filters in every frame for the (a) centre, (b) front left, (c) front right, (d) side left and (e) side right channels when using the DEA tap selection scheme for $Q = 0.2$ for a 30s segment of a 5-channel concert signal.

For a 5-channel concert signal, Figure 2.12 shows the ERLE curves obtained when the MAEC filters are updated using the different tap selection schemes for $Q = 0.2$ and for full filter update ($Q = 1$) for a period of 30s. For this signal, we find that even when only 20% of all filter taps are updated in every frame, both the 3DM scheme and the proposed DEA scheme perform almost identically to full filter update in terms of ERLE, with less than 1 dB deterioration, while the FEA scheme leads to about 2-4 dB deterioration in terms of ERLE. The SPU scheme again performs significantly worse, yielding about 10-12 dB deterioration in ERLE. It can be seen that around the 12s mark, all schemes witness a sudden drop in ERLE. This is because the tap-input covariance matrix becomes ill-conditioned, leading to an increase in misalignment. However, it can also be observed that even though the FEA and DEA schemes have not been designed to tackle the misalignment problem, they do not deteriorate the problem further.

Additionally, Figure 2.13 shows the number of taps $\mathcal{L}_r(k,\ell)$ updated in the different sub-filters in every frame using the DEA scheme for $Q = 0.2$. It can be observed that the sub-filters in each channel get a small number of taps selected in every frame, where the number of taps updated across subbands depends on the spectral content present in each channel. As the centre channel for this signal consists of only speech, the tap allocation for the centre channel strongly resembles the spectrogram of a speech signal. As the surround-sound channels are mainly dominated by background

Table 2.1: Computational Effort: Number of operations per frame for implementing the different tap selection schemes and for updating the MAEC filters using the PUNLMS algorithm

Ops	3DM	SPU	FEA	DEA	PUNLMS
# Adds	KR	$3KR$	0	$6KR+1$	$4QN+3KR-K$
# Mults	0	$2KR$	KR	$3KR+2$	$4QN+2KR+3K$
# Divs	0	0	0	2	K
# Comps	$N\log_2 N$	$KR\log_2(KR)$	$KR(2\log_2 L+2)$	$KR(2\log_2 L+3)+1$	0

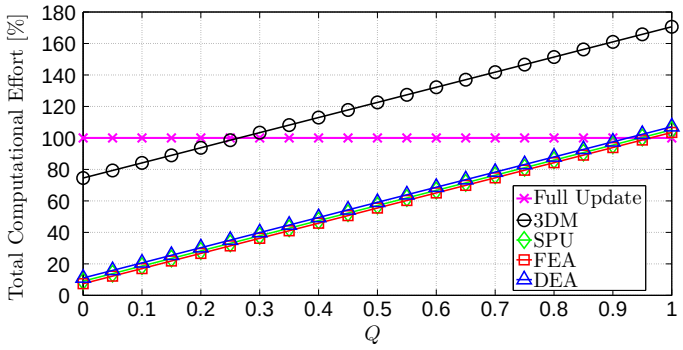

Fig. 2.14: Total computational effort required per frame for implementing the different tap selection schemes and for updating the MAEC filters using the PUNLMS algorithm as a function of Q. The numbers have been computed for $K = 257$, $R = 5$ and $L = 20$ and have been plotted as a percentage of the effort required for full filter update.

score and low-frequency crowd noise but also contain some speech, this is reflected in how taps are allocated in the surround-sound channels.

2.6 Computational effort

When compared to full filter update, implementing a tap selection scheme requires some computational overhead, but still may result in significant savings when updating the MAEC filters, as only a fraction Q of the total N filter taps are updated in every frame. The computational effort per frame for implementing the different tap selection schemes and for updating the MAEC filters using the PUNLMS algorithm is given in Table 2.1. The computations have been divided into four categories, namely the number of additions (# Adds), multiplications (# Mults), divisions (# Divs) and comparisons (# Comps). Please note that all complex operations have been converted into an equivalent number of real operations, e.g. 1 complex multiplication has been counted as 4 real multiplications and 2 real additions.

Figure 2.14 is an exemplary figure depicting the total computational effort required per frame for implementing tap selection and partial filter update for different values of Q. The numbers have been computed for $K = 257$, $R = 5$ and $L = 20$, and by assuming that the comparison, multiplication and division operations are 1, 4 and 15 times as computationally expensive as an addition operation, respectively. The numbers have been plotted as a percentage of the computational effort required for full filter update, i.e., the effort required for updating the MAEC filters using the PUNLMS algorithm with $Q = 1$. For these assumed settings, it can be observed that the total computational effort for the 3DM, SPU, FEA and DEA schemes is smaller than full filter update for $Q < 0.27$, $Q < 0.95$, $Q < 0.96$ and $Q < 0.93$, respectively. Hence, the SPU, FEA and DEA schemes are almost always cheaper than full filter update. When only 20% of the MAEC filter taps are updated in every frame ($Q = 0.2$), the 3DM scheme requires 94%, while the SPU, FEA and DEA schemes require about 28% of the total computational effort required for full filter update. Using the SPU and DEA schemes results in slightly larger computational effort as compared to the FEA scheme due to the additional overhead required for computing $\eta_r(k,\ell)$ in Equation (2.18) and $\psi_r^D(k,\ell)$ in Equation (2.35), respectively.

2.7 Conclusions

In this paper, different tap selection schemes for constrained partial updates of subband MAEC filters have been compared. Real-world multichannel signals have been analyzed and shown to be sparse across subbands (spectrally), channels (spatially) and frames (temporally). This sparsity is then exploited by different tap selection schemes for updating the MAEC filters. The MAEC system consists of a dedicated subband AEC filter for each loudspeaker channel, with each filter composed of multiple sub-filters, i.e., one sub-filter per subband per channel. The first tap selection scheme considered applied the well-known M-Max criterion on the multichannel input spectra across all three dimensions, and is hence called the 3DM scheme. This scheme jointly exploits the spectral, spatial and temporal sparsity in the input signals but typically results in some sub-filters having no taps updated. In order to avoid this problem, two new schemes have been presented which perform tap selection by applying the M-Max criterion only across filter length (and thereby exploit temporal sparsity for updating each sub-filter) and do not completely ignore the sub-filters with the smallest magnitude tap-inputs. The FEA scheme allocates a fixed number of taps to be updated in each sub-filter per frame, while the proposed DEA scheme exploits the joint spectro-spatial sparsity present in the input signals for dynamically allocating the number of taps to be updated in the different sub-filters. The new tap selection schemes have been compared to the state-of-the-art SPU tap selection scheme in the subband domain, which displays similar properties to the 3DM scheme. The proposed DEA scheme is designed such that it selects more taps in the sub-filters with larger magnitude tap-inputs (like the 3DM and SPU schemes) while not completely ignoring the sub-filters with smaller magnitude tap-inputs (like the FEA scheme). Simulation results for speech and music signals showed that in terms of ERLE and convergence speed, the 3DM and DEA schemes

achieved almost identical echo cancellation performance compared to full filter update even when only 20% of the MAEC filter taps were updated in every frame, while the FEA and SPU schemes performed worse (about 2-4 dB and 10-12 dB deterioration in ERLE, respectively). The SPU, FEA and DEA tap selection schemes have a reduced computational cost compared to full filter update, while the 3DM scheme does not necessarily lead to reduction in computational complexity. Hence, in conclusion, the proposed DEA tap selection scheme yields almost identical echo cancellation performance compared to updating all filter taps at a significantly reduced computational cost.

3

ONLINE ESTIMATION OF REVERBERATION PARAMETERS FOR LATE RESIDUAL ECHO SUPPRESSION

3.1 Abstract

In hands-free telephony and other distant-talk applications, often a short AEC filter is used to achieve fast convergence at low computational cost. As a result, a significant amount of late residual echo (LRE) may remain, especially in highly reverberant environments. This LRE can be suppressed using a postfilter in the subband domain, which requires an estimate of the power spectral density (PSD) of the LRE. To estimate the LRE PSD, an exponentially decaying model with frequency-dependent reverberation scaling and decay parameters has frequently been assumed. State-of-the-art methods estimate both reverberation parameters independently of each other, either in offline or in online mode. In this article, we propose two signal-based methods (i.e., output error and equation error) to jointly estimate both reverberation parameters in online mode. The estimated parameters are then used to generate an estimate for the LRE PSD, which is fed into a postfilter for the purpose of late residual echo suppression. We derive several gradient-descent-based algorithms to simultaneously update both reverberation parameters, minimizing either the mean squared error or the mean squared log error cost function. The proposed methods are compared with state-of-the-art methods in terms of the accuracy of the estimated reverberation parameters and the corresponding LRE PSD estimate. Extensive simulation results using both artificial as well as measured room impulse responses show that the proposed output error method with mean squared log error minimization outperforms state-of-the-art methods in all considered scenarios.

3.2 Introduction

Hands-free telephony and other distant-talk applications, such as voice-controlled multimedia devices, are often used in large reverberant rooms, where the distance between the desired (near-end) speaker and the microphone may be quite large. Due to the acoustic coupling between the loudspeaker and the microphone, the microphone signal is typically degraded by the acoustic echo of the far-end signal,

which may significantly reduce the quality and/or the intelligibility of the near-end speaker. Acoustic echo cancellation (AEC) [1] is a key technology used in such scenarios, aimed at canceling the echo from the microphone signal. An AEC system typically consists of an adaptive filter [10,15] which estimates the acoustic echo path, i.e., the room impulse response (RIR) between the loudspeaker and the microphone. The adaptive filter is used to generate an estimate of the acoustic echo signal, which is subsequently subtracted from the microphone signal. The resulting signal is referred to as the AEC error signal and is composed of near-end speech, background noise and usually some residual echo, as the AEC filter is unable to completely accurately estimate the RIR in practice (filter misalignment). When deploying an AEC system in a room with a large reverberation time (T_{60}), a large filter length needs to be used in order to achieve good echo cancellation performance. However, using a long filter results in large computational cost for updating the filter and may also lead to slow filter convergence [10, 15]. Hence, aiming at achieving fast filter convergence at low computational cost, in practice often a short AEC filter is used, which however results in a large amount of late residual echo (LRE).

In practice, a postfilter is often used in addition to the AEC filter, aimed at suppressing the residual echo and background noise while not distorting the near-end speech signal. Although multi-frame postfilters have been proposed [68], most postfilters are single-tap real-valued gains [14, 21, 22, 62–64, 66, 67]. To design the postfilter in the subband domain, an accurate estimate of the power spectral density (PSD) of the residual echo and background noise signals is required. A simple but frequently used method to estimate the PSD of the residual echo signal is to apply a coupling factor to the far-end signal PSD, where the coupling factor is estimated during periods of near-end speech absence [1]. However, since this method does not take into account any temporal context and is unable to model the LRE PSD accurately, its performance is quite poor, especially when using a short AEC filter. Hence, several other LRE PSD estimators have been proposed which are based on the statistical reverberation model proposed in [28, 72], which assumes that the late reverberant part of a RIR decays exponentially at a rate proportional to the T_{60}. These PSD estimators require estimates of two parameters: the reverberation decay parameter (corresponding to the T_{60}) and the reverberation scaling parameter (corresponding to the initial power of the LRE).

To estimate both reverberation parameters, channel-based as well as signal-based methods have been proposed. *Channel-based* methods [13,14] estimate the reverberation parameters using the coefficients of the converged AEC filter, either assuming frequency-dependent [14] or frequency-independent parameters [13]. Channel-based methods are only effective if relatively long AEC filters are used, which are able to capture the decay of the late reverberant part of the RIR. *Signal-based* methods, on the other hand, estimate the reverberation parameters directly from the far-end and the residual echo signals [23, 24, 77]. In [23], a signal-based method was proposed to estimate both reverberation parameters in *offline* mode (i.e., batch processing). The reverberation scaling parameter was estimated by minimizing the mean squared error (MSE) cost function, while the reverberation decay parameter was estimated by minimizing the mean squared log error (MSLE) cost function. In [24], a pure

acoustic echo suppression system, i.e., without an AEC filter, was considered and a recursive estimator for the (residual) echo PSD was used. A signal-based method exploiting higher-order-statistics was proposed to estimate the initial power of the (residual) echo and the reverberation decay parameter independently of each other in *online mode*. Using the recursive estimator for the LRE PSD in [14], in [77] we proposed two signal-based methods, namely an output error and an equation error method, to *jointly* estimate both reverberation parameters in offline mode. These methods, which were originally proposed to estimate the coefficients of generic IIR filters in the time-domain [15, 75, 76], were applied on PSDs to jointly estimate both reverberation parameters by minimizing either the MSE or the MSLE cost function.

Based on the work in [77], in this paper we propose methods to *jointly* estimate both reverberation parameters in *online* mode. The estimated parameters are then used to generate an estimate for the LRE PSD, which is fed into a postfilter for the purpose of late residual echo suppression. We derive several gradient-descent-based algorithms to simultaneously update both parameters, minimizing either the MSE or the MSLE cost function. In particular, we propose to use the recursive prediction error (RPE) and pseudo-linear regression (PLR) algorithms, which were derived for time-domain recursive systems [75], to update the parameters for the output error method. The different signal-based methods (output/equation error), algorithms (RPE/PLR) and cost functions (MSE/MSLE) are compared with state-of-the-art signal-based methods [23, 24] in terms of accuracy of the reverberation parameter estimates and the corresponding LRE PSD estimate, and in terms of the resulting residual echo suppression and near-end speech distortion.

The paper is organized as follows. The signal model as well as some basic AEC and postfiltering principles are presented in Section 3.3. The recursive estimator for the LRE PSD, the different proposed signal-based parameter estimation methods and the gradient-descent-based algorithms to simultaneously update both parameters are presented in Sections 3.4, 3.5 and 3.6, respectively. Section 3.7 presents the simulation results comparing the performance of the proposed methods with state-of-the-art methods using both artificially generated as well as measured RIRs.

3.3 Signal model and AEC system

Fig. 3.1 shows a loudspeaker-enclosure-microphone (LEM) system with the far-end signal x, the acoustic echo signal d, the near-end speech signal s, the background noise signal v and the microphone signal y. The RIR characterizing the acoustic echo path between the loudspeaker and the microphone is denoted as h and assumed to be time-invariant and of length N_h. The microphone signal at discrete-time sample n is given as:

$$y(n) = s(n) + v(n) + \underbrace{\sum_{i=0}^{N_h-1} h(i) \cdot x(n-i)}_{d(n)} . \tag{3.1}$$

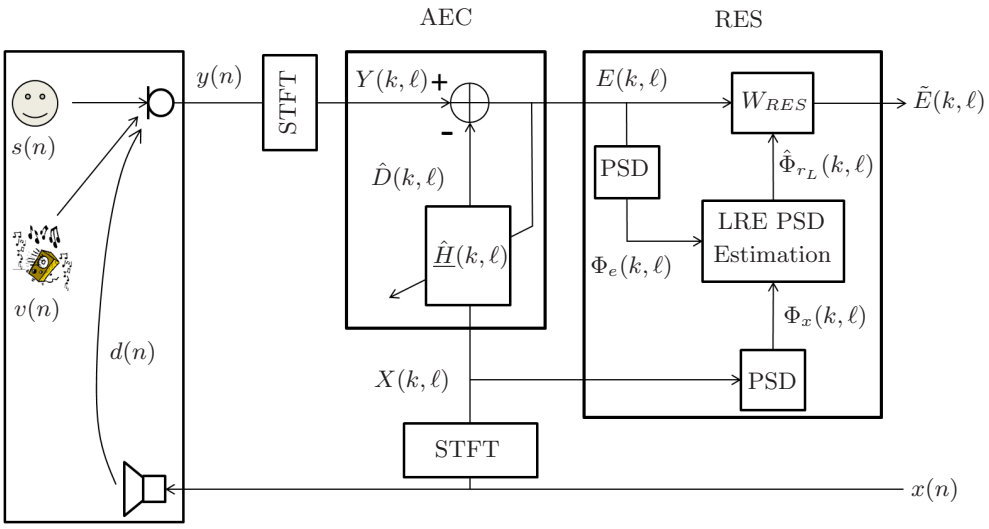

Fig. 3.1: Acoustic echo cancellation (AEC) and residual echo suppression (RES) systems.

For the subband processing, a fast Fourier transform (FFT) filterbank of order N_{FFT} is used to transform the (windowed) time-domain signals into the short-time Fourier transform (STFT) domain, with the total number of subbands given by $K = \frac{N_{\text{FFT}}}{2} + 1$. The complex-valued STFT coefficients of the far-end signal x in subband k and frame ℓ are computed as:

$$X(k,\ell) = \sum_{m=0}^{N_{\text{FFT}}-1} x(\ell \cdot F + m) \cdot W_{\text{ana}}(m) \cdot e^{-j\frac{2\pi}{N_{\text{FFT}}}km}, \tag{3.2}$$

where $j = \sqrt{-1}$, F denotes the frameshift and W_{ana} denotes the analysis window. Similarly to (3.2), the STFT coefficients of $s(n)$, $v(n)$, $d(n)$ and $y(n)$ are denoted as $S(k,\ell)$, $V(k,\ell)$, $D(k,\ell)$ and $Y(k,\ell)$, respectively.

The complete AEC system consists of two components: an (adaptive) AEC filter estimating the echo path and a residual echo suppression (RES) postfilter. Both components will be explained in more detail in the following subsections.

3.3.1 *Acoustic echo cancellation*

To cancel the acoustic echo signal from the microphone signal, we consider a G-tap subband AEC filter $\underline{\hat{H}}$. The acoustic echo estimate is given as:

$$\hat{D}(k,\ell) = \underline{X}^H(k,\ell)\ \underline{\hat{H}}(k), \tag{3.3}$$

with

$$\underline{X}(k,\ell) = \left[\ X(k,\ell) \quad \ldots \quad X(k,\ell-G+1)\ \right]^T \tag{3.4}$$

the G-dimensional tap-input vector to the AEC filter $\underline{\hat{H}}$:

$$\underline{\hat{H}}(k) = \begin{bmatrix} \hat{H}_1(k) & \dots & \hat{H}_G(k) \end{bmatrix}^T, \tag{3.5}$$

where $\cdot^H$ denotes the Hermitian operator and $\cdot^T$ denotes the transpose operator.

The signal obtained after the acoustic echo estimate is subtracted from the microphone signal is referred to as the AEC error signal:

$$\begin{aligned} E(k,\ell) &= Y(k,\ell) - \hat{D}(k,\ell) \\ &= S(k,\ell) + V(k,\ell) + \underbrace{\Big(D(k,\ell) - \hat{D}(k,\ell)\Big)}_{R(k,\ell)}, \end{aligned} \tag{3.6}$$

where R denotes the residual echo signal, which consists of the early residual echo signal R_E (due to filter misalignment) and the LRE signal R_L (due to the limited length of the AEC filter). In this paper, the filter length G is chosen so as to cover the direct path and the early reflections of the RIR h, i.e., $G = \lfloor \frac{N}{F} \rfloor$, where N corresponds to the length of the direct path and early reflections in samples. This means that the LRE signal R_L is assumed to contain only late reverberation. Additionally, we assume no filter misalignment, i.e., $R_E = 0$, such that the residual echo signal only consists of the late residual echo signal, i.e., $R = R_L$.

3.3.2 *Residual echo suppression*

Residual echo suppression can be performed in the subband domain by applying a real-valued gain W_{RES} to the AEC error signal E, as shown in Fig. 3.1. A frequently used gain is the Wiener filter [1], which is derived by assuming that the signals S, R_L and V are independent stationary stochastic processes, leading to:

$$W_{\text{RES}}(k,\ell) = 1 - \frac{\lambda_{r_L}(k,\ell) + \lambda_v(k,\ell)}{\lambda_e(k,\ell)}. \tag{3.7}$$

Here, λ_{r_L}, λ_v and λ_e denote the PSDs of the LRE, the background noise and the AEC error signals, respectively, defined as $\lambda_{r_L}(k,\ell) = \mathcal{E}\{|R_L(k,\ell)|^2\}$, $\lambda_v(k,\ell) = \mathcal{E}\{|V(k,\ell)|^2\}$, and $\lambda_e(k,\ell) = \mathcal{E}\{|E(k,\ell)|^2\}$, where $\mathcal{E}\{\cdot\}$ denotes the statistical expectation operator. In practice, the statistical expectation operator is approximated by temporal averaging (assuming ergodicity), e.g.:

$$\Phi_e(k,\ell) = \alpha \cdot \Phi_e(k,\ell-1) + (1-\alpha) \cdot |E(k,\ell)|^2, \tag{3.8}$$

where Φ_e is an approximation of the PSD λ_e and α denotes the smoothing factor. The quantities Φ_{r_L}, Φ_v and Φ_x are defined similarly as in (3.8) and are approximations of λ_{r_L}, λ_v and λ_x, respectively. Please note that for an unobservable signal such as r_L, the quantity Φ_{r_L} itself needs to be estimated, with the estimate denoted as $\hat{\Phi}_{r_L}$. In the remainder of the paper, we will use the term *true PSD* to refer to λ_a,

$a \in \{e, r_L, v, x\}$, the term *PSD* to refer to its approximation Φ_a, $a \in \{e, r_L, v, x\}$ and the term *PSD estimate* to refer to its estimate for an unobservable signal $\hat{\Phi}_a$, $a \in \{r_L, v\}$.

In order to control the aggressiveness of the residual echo suppression, we use the following gain for the RES postfilter:

$$W_{\mathrm{RES}}(k,\ell) = \max\left\{1 - \beta \cdot \left(\frac{\hat{\Phi}_{r_L}(k,\ell) + \hat{\Phi}_v(k,\ell)}{\Phi_e(k,\ell)}\right),\ \gamma\right\} \tag{3.9}$$

with over-estimation factor β and spectral floor γ. While the AEC error PSD Φ_e is directly observable, the LRE PSD Φ_{r_L} and the background noise PSD Φ_v need to be estimated. Many approaches have been proposed in literature for estimating the background noise PSD [89–91]. In this paper, we assume that the background noise is stationary and its PSD is known.

The processed AEC error signal is given as:

$$\tilde{E}(k,\ell) = W_{\mathrm{RES}}(k,\ell) \cdot E(k,\ell), \tag{3.10}$$

which can be expressed as the sum of its individual components in a similar way to (3.6):

$$\tilde{E}(k,\ell) = \tilde{S}(k,\ell) + \tilde{V}(k,\ell) + \tilde{R}_L(k,\ell), \tag{3.11}$$

where $\tilde{S}$, $\tilde{V}$ and $\tilde{R}_L$ are obtained by multiplying S, V and R_L with the RES postfilter, similarly to (3.10). For the purpose of evaluation, these processed signals are then synthesized to the time-domain using inverse STFT and overlap-add processing, yielding the time-domain signals $\tilde{e}(n)$, $\tilde{s}(n)$, $\tilde{v}(n)$ and $\tilde{r}_L(n)$, respectively.

3.4 Model for LRE PSD

In [28], an exponentially decaying model for the late reverberant part of a RIR was proposed when the source-microphone distance is larger than the critical distance, defined as the distance where the energy of the direct sound is equal to the energy of all reflections [25]. According to this model, the late reverberant part of a RIR can be described as a realization of a stochastic process:

$$h(i) = w_L(i) \cdot e^{-\rho(i-N)}, \quad N \le i < N_h, \tag{3.12}$$

where N_h denotes the total length of the RIR in samples, w_L is a zero-mean white Gaussian noise process with variance σ_L^2 and ρ denotes the decay rate. The decay rate is related to the T_{60} as:

$$\rho = \frac{3 \cdot \ln 10}{f_s \cdot T_{60}}, \tag{3.13}$$

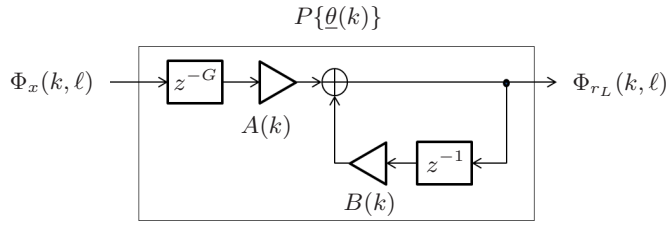

Fig. 3.2: Model for LRE PSD Φ_{r_L} as a function of far-end signal PSD Φ_x.

where f_s denotes the sampling frequency in Hz. Although in (3.13) it is assumed that the T_{60} is frequency-independent, it should be noted that in practice the T_{60} (and hence the decay rate ρ) is frequency-dependent [25].

As mentioned in Section 3.3.1, we assume that the AEC filter is able to cancel the direct sound component and the early reflections, such that the LRE signal R_L contains only late reverberation. Based on the RIR model in (3.12), a recursive expression for λ_{r_L} can be derived (see Appendix A.1), i.e.:

$$\lambda_{r_L}(k,\ell) = A \cdot \lambda_x(k,\ell-G) + B \cdot \lambda_{r_L}(k,\ell-1), \tag{3.14}$$

where A denotes the reverberation scaling parameter and B denotes the reverberation decay parameter. These parameters are related to the parameters σ_L^2 and ρ of the RIR model in (3.12) as (see Appendix A.1):

$$A = \sigma_L^2 \cdot \left(\frac{1-e^{-2\rho F}}{1-e^{-2\rho}}\right), \tag{3.15}$$

$$B = e^{-2\rho F}. \tag{3.16}$$

In this paper, we assume the reverberation parameters to be *frequency-dependent*, such that similarly to (3.14), a recursive expression for Φ_{r_L} using frequency-dependent parameters can be obtained as in [14]:

$$\boxed{\Phi_{r_L}(k,\ell) = A(k) \cdot \Phi_x(k,\ell-G) + B(k) \cdot \Phi_{r_L}(k,\ell-1),} \tag{3.17}$$

with the parameters $A(k)$ and $B(k)$ given as:

$$A(k) = \sigma_L^2(k) \cdot \left(\frac{1-e^{-2\rho(k)F}}{1-e^{-2\rho(k)}}\right), \tag{3.18}$$

$$B(k) = e^{-2\rho(k)F}. \tag{3.19}$$

The expression in (3.17) relating the LRE PSD Φ_{r_L} to the far-end signal PSD Φ_x is illustrated in Fig. 3.2 using the IIR filter $P\{\underline{\theta}(k)\}$, where

$$\underline{\theta}(k) = \begin{bmatrix} A(k) & B(k) \end{bmatrix}^T. \tag{3.20}$$

In the next section, we will present different methods to estimate $\underline{\theta}(k)$. It should be noted that $\underline{\theta}(k)$ is estimated during periods of near-end speech absence and subsequently used to estimate the LRE PSD during periods of double-talk.

3.5 Parameter estimation methods

Several methods have been proposed in literature to estimate both reverberation parameters A and B independently of each other. In [14], a channel-based method was proposed using the converged AEC filter coefficients. In [23], a signal-based method was proposed in offline mode (i.e., batch processing), where the parameter A was estimated by minimizing an MSE cost function and the parameter B was estimated by minimizing an MSLE cost function. In [24], an acoustic echo suppression setup without an AEC filter (i.e., $G = 0$) was considered and a signal-based method based on higher-order statistics was proposed to estimate both parameters in online mode. For the purpose of fair comparison, we consider a slightly modified version of the method in [24] in order to estimate the LRE PSD for our considered setup and compare this method with our proposed parameter estimation methods (see Section 3.7). Since we assume a perfect AEC filter (see Section 3.3.1), this modification simply corresponds to inserting a delay of G frames in the original method in [24] (details presented in Appendix A.2).

To *jointly* estimate the parameters of generic IIR filters in the time-domain, several signal-based methods have been proposed [15, 75, 76, 92–94], either based on the output error (OE) or the equation error (EE). In [77], we applied the OE and EE methods on PSDs to jointly estimate both reverberation parameters in offline mode (i.e., batch processing), minimizing either the MSE or the MSLE cost function. Simulation results showed that the most accurate estimates for the reverberation decay parameter B and the LRE PSD Φ_{r_L} were obtained using the OE method minimizing the MSLE cost function, while the most accurate estimates for the reverberation scaling parameter A were obtained using either the OE or the EE method minimizing the MSE cost function.

Based on the offline methods from [77], in this paper we investigate the OE and EE methods in *online* mode to jointly estimate both reverberation parameters A and B during periods of near-end speech absence, where the parameters are simultaneously updated in each frame using a gradient-descent-based algorithm (see Section 3.6). The estimated parameters $\hat{\underline{\theta}}(k)$ are then fed into the IIR filter $P\{\hat{\underline{\theta}}(k,\ell)\}$ to estimate the LRE PSD (also during double-talk), as illustrated in Fig. 3.3:

$$\boxed{\hat{\Phi}_{r_L}(k,\ell) = \hat{A}(k,\ell) \cdot \Phi_x(k,\ell-G) + \hat{B}(k,\ell) \cdot \hat{\Phi}_{r_L}(k,\ell-1).} \tag{3.21}$$

In the following subsections we will discuss the OE and EE methods to estimate the reverberation parameters $\hat{\underline{\theta}}(k)$.

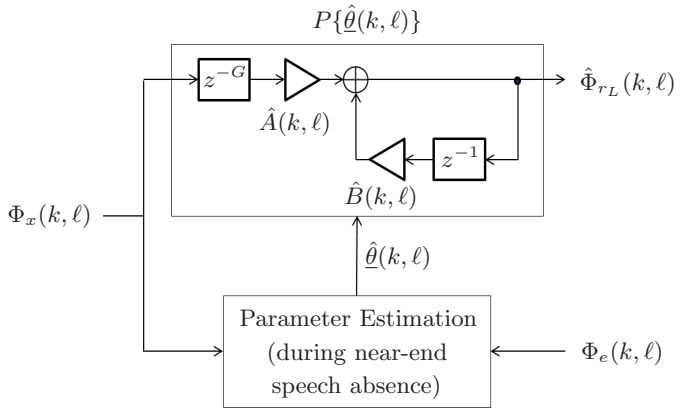

Fig. 3.3: The LRE PSD estimate $\hat{\Phi}_{r_L}$ is computed using the far-end signal PSD Φ_x, with the parameters $\underline{\hat{\theta}}(k,\ell)$ estimated during near-end speech absence.

3.5.1 *Output error method*

The OE method is a well-known method used for parameter estimation of linear recursive systems in a variety of applications. The OE method is characterized by the following *recursive* difference equation (where the superscript $^{\mathrm{O}}$ denotes the OE method):

$$\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell) = \hat{A}^{\mathrm{O}}(k,\ell)\cdot\Phi_x(k,\ell-G) + \hat{B}^{\mathrm{O}}(k,\ell)\cdot\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1), \tag{3.22}$$

with the corresponding IIR filter structure illustrated in Fig. 3.4. Here, $\hat{\Phi}^{\mathrm{O}}_{r_L}$ denotes the OE PSD estimate and

$$\underline{\hat{\theta}}^{\mathrm{O}}(k,\ell) = \begin{bmatrix} \hat{A}^{\mathrm{O}}(k,\ell) & \hat{B}^{\mathrm{O}}(k,\ell) \end{bmatrix}^T \tag{3.23}$$

denotes the reverberation parameters estimated using the OE method, which are fed into (3.21) to generate the LRE PSD estimate $\hat{\Phi}_{r_L}$. Please note that (3.22) has the same recursive structure as (3.21), such that $\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell) = \hat{\Phi}_{r_L}(k,\ell)$. From (3.22), it can be observed that the OE PSD estimate in the current frame $\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell)$ not only depends on the parameter estimates in the current frame $\underline{\hat{\theta}}^{\mathrm{O}}(k,\ell)$, but also on the OE PSD estimate in the previous frame $\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)$, which itself depends on the parameter estimates in the previous frame $\underline{\hat{\theta}}^{\mathrm{O}}(k,\ell-1)$, and so on. Thus, $\hat{\Phi}^{\mathrm{O}}_{r_L}$ is a *non-linear* function of $\underline{\hat{\theta}}^{\mathrm{O}}$, where the current OE PSD estimate depends on the parameter estimates in all previous frames.

The output error is obtained by subtracting the output in (3.22) from the target PSD Φ_{r_L}:

$$Q^{\mathrm{O}}(k,\ell) = \Phi_{r_L}(k,\ell) - \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell). \tag{3.24}$$

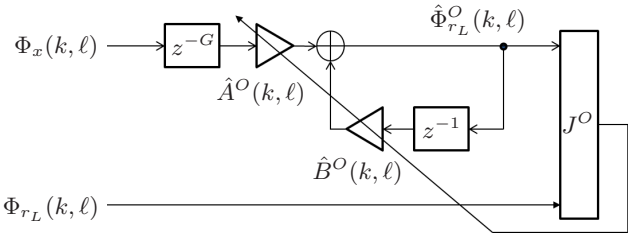

Fig. 3.4: Parameter estimation using the output error method by minimizing the cost function J^{O}.

Similarly, the output log error is given as:

$$Q^{\mathrm{O}}_{\ln}(k,\ell) = \ln\ \Phi_{r_L}(k,\ell) - \ln\ \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell) = \ln\left(\frac{\Phi_{r_L}(k,\ell)}{\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell)}\right). \tag{3.25}$$

To compute the parameter estimates, we will consider minimizing either the MSE or the MSLE cost function:

$$\mathcal{J}^{\mathrm{O}}_{\mathrm{MSE}}\ \left(\hat{A}^{\mathrm{O}}(k,\ell), \hat{B}^{\mathrm{O}}(k,\ell)\right) = \mathcal{E}\left\{\left[Q^{\mathrm{O}}(k,\ell)\right]^2\right\}, \tag{3.26}$$

$$\mathcal{J}^{\mathrm{O}}_{\mathrm{MSLE}}\ \left(\ln\hat{A}^{\mathrm{O}}(k,\ell), \ln\hat{B}^{\mathrm{O}}(k,\ell)\right) = \mathcal{E}\left\{\left[Q^{\mathrm{O}}_{\ln}(k,\ell)\right]^2\right\}. \tag{3.27}$$

To update the parameters in every frame using a gradient-descent-based algorithm (see Section 3.6), these cost functions will be approximated by their instantaneous values:

$$J^{\mathrm{O}}_{\mathrm{MSE}}\ \left(\hat{A}^{\mathrm{O}}(k,\ell), \hat{B}^{\mathrm{O}}(k,\ell)\right) = \left[Q^{\mathrm{O}}(k,\ell)\right]^2 = \left[\Phi_{r_L}(k,\ell) - \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell)\right]^2, \tag{3.28}$$

$$J^{\mathrm{O}}_{\mathrm{MSLE}}\ \left(\ln\hat{A}^{\mathrm{O}}(k,\ell), \ln\hat{B}^{\mathrm{O}}(k,\ell)\right) = \left[Q^{\mathrm{O}}_{\ln}(k,\ell)\right]^2 = \left[\ln\left(\frac{\Phi_{r_L}(k,\ell)}{\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell)}\right)\right]^2. \tag{3.29}$$

As $\hat{\Phi}^{\mathrm{O}}_{r_L}$ is a non-linear function of the parameters $\underline{\hat{\theta}}^{\mathrm{O}}$, the cost functions $J^{\mathrm{O}}_{\mathrm{MSE}}$ and $J^{\mathrm{O}}_{\mathrm{MSLE}}$ are not quadratic in the parameters and may exhibit multiple local minima [75, 94–97]. This may result in gradient-descent-based algorithms converging to a local minimum, thereby yielding sub-optimal and inaccurate parameter estimates, with the initial value of $\underline{\hat{\theta}}^{\mathrm{O}}$ also influencing to which minimum the algorithms converge. This is a typical problem when using adaptive IIR filters for identifying recursive systems [75].

3.5.2 *Equation error method*

In order to avoid the local minima problem associated with the OE method, the EE method has often been employed for parameter estimation of linear recursive

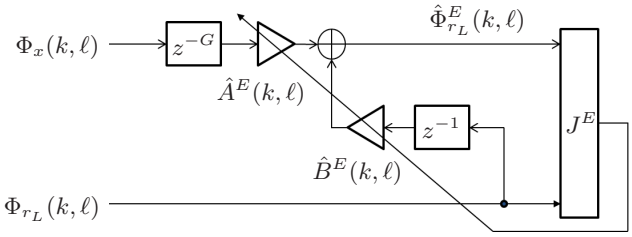

Fig. 3.5: Parameter estimation using the equation error method by minimizing the cost function J^{E}.

systems [75, 76]. The EE method differs from the OE method by using the delayed target PSD $\Phi_{r_L}(k,\ell-1)$ instead of the delayed PSD estimate $\hat{\Phi}_{r_L}(k,\ell-1)$ for computing the current PSD estimate, thereby breaking the recursive structure. The EE method is characterized by the following *non-recursive* difference equation (where the superscript $^{\mathrm{E}}$ denotes the EE method):

$$\hat{\Phi}^{\mathrm{E}}_{r_L}(k,\ell) = \hat{A}^{\mathrm{E}}(k,\ell)\cdot\Phi_x(k,\ell-G) + \hat{B}^{\mathrm{E}}(k,\ell)\cdot\Phi_{r_L}(k,\ell-1), \tag{3.30}$$

with the corresponding non-recursive filter structure illustrated in Fig. 3.5. Here, $\hat{\Phi}^{\mathrm{E}}_{r_L}$ denotes the EE PSD estimate and

$$\underline{\hat{\theta}}^{\mathrm{E}}(k,\ell) = \begin{bmatrix} \hat{A}^{\mathrm{E}}(k,\ell) & \hat{B}^{\mathrm{E}}(k,\ell) \end{bmatrix}^T \tag{3.31}$$

denotes the reverberation parameters estimated using the EE method, which are fed into (3.21) to generate the LRE PSD estimate $\hat{\Phi}_{r_L}$. As a result, the PSD estimate $\hat{\Phi}^{\mathrm{E}}_{r_L}$ is a *linear* function of $\underline{\hat{\theta}}^{\mathrm{E}}$. Please note that using $\Phi_{r_L}(k,\ell-1)$ instead of $\hat{\Phi}_{r_L}(k,\ell-1)$ in (3.30) is an approximation, such that unlike the OE method, the EE PSD estimate $\hat{\Phi}^{\mathrm{E}}_{r_L}$ is not equal to the LRE PSD estimate $\hat{\Phi}_{r_L}$.

Similarly to (3.24), the equation error is given as:

$$Q^{\mathrm{E}}(k,\ell) = \Phi_{r_L}(k,\ell) - \hat{\Phi}^{\mathrm{E}}_{r_L}(k,\ell), \tag{3.32}$$

and similarly to (3.25), the equation log error is given as:

$$Q^{\mathrm{E}}_{\ln}(k,\ell) = \ln\ \Phi_{r_L}(k,\ell) - \ln\ \hat{\Phi}^{\mathrm{E}}_{r_L}(k,\ell) = \ln\left(\frac{\Phi_{r_L}(k,\ell)}{\hat{\Phi}^{\mathrm{E}}_{r_L}(k,\ell)}\right). \tag{3.33}$$

Contrary to the cost functions $J^{\mathrm{O}}_{\mathrm{MSE}}$ and $J^{\mathrm{O}}_{\mathrm{MSLE}}$, the cost functions $J^{\mathrm{E}}_{\mathrm{MSE}}$ and $J^{\mathrm{E}}_{\mathrm{MSLE}}$ (defined similarly to (3.28) and (3.29), respectively) are quadratic in the parameters, hence exhibiting a single global minimum and no local minima [75, 76]. This makes the EE method particularly attractive for use in practical applications, as the corresponding adaptive algorithms typically have fast convergence and converge to a global minimum. However, it has been shown in [75] that the EE method yields biased solutions in the presence of additive noise, where the bias is proportional to

the amount of noise. Additionally, as Φ_x and Φ_{r_L} are approximations of the true PSDs λ_x and λ_{r_L} (see Section 3.3.2), these approximations introduce additional noise to the system. This results in the EE method yielding biased solutions even in the absence of additive noise, as was observed in [77] when using the EE method for reverberation parameter estimation in offline mode. In this paper, we investigate how accurately the EE method estimates the reverberation parameters in online mode.

3.6 Gradient-descent-based algorithms

In this section, we derive gradient-descent-based algorithms to update the reverberation parameters $\underline{\theta}(k)$ in every frame for the OE and EE estimation methods, either minimizing the MSE or the MSLE cost function.

For both estimation methods, the gradient-descent update rule for the MSE cost function is given as:

$$\boxed{\underline{\hat{\theta}}^{\mathrm{I}}(k,\ell+1) = \underline{\hat{\theta}}^{\mathrm{I}}(k,\ell) - \frac{\underline{\Gamma}}{2} \odot \underline{\nabla}^{\mathrm{I}}_{\mathrm{MSE}}(k,\ell),} \tag{3.34}$$

where $\mathrm{I} \in \{\mathrm{O},\mathrm{E}\}$ denotes the used estimation method, $\odot$ denotes element-wise multiplication, $\underline{\Gamma} = \begin{bmatrix} \mu_A & \mu_B \end{bmatrix}^T$ denotes the (fixed) step-sizes to update both parameters, and

$$\underline{\nabla}^{\mathrm{I}}_{\mathrm{MSE}}(k,\ell) = \begin{bmatrix} \frac{\partial J^{\mathrm{I}}_{\mathrm{MSE}}\left(\hat{A}^{\mathrm{I}}(k,\ell),\hat{B}^{\mathrm{I}}(k,\ell)\right)}{\partial \hat{A}^{\mathrm{I}}(k,\ell)} & \frac{\partial J^{\mathrm{I}}_{\mathrm{MSE}}\left(\hat{A}^{\mathrm{I}}(k,\ell),\hat{B}^{\mathrm{I}}(k,\ell)\right)}{\partial \hat{B}^{\mathrm{I}}(k,\ell)} \end{bmatrix}^T \tag{3.35}$$

denotes the gradient of the MSE cost function. Using (3.28), the partial derivatives of the MSE cost function with respect to the reverberation scaling and decay parameter estimates are equal to:

$$\frac{\partial J^{\mathrm{I}}_{\mathrm{MSE}}\left(\hat{A}^{\mathrm{I}}(k,\ell),\hat{B}^{\mathrm{I}}(k,\ell)\right)}{\partial \hat{A}^{\mathrm{I}}(k,\ell)} = -2 \cdot Q^{\mathrm{I}}(k,\ell) \cdot \frac{\partial \hat{\Phi}^{\mathrm{I}}_{r_L}(k,\ell)}{\partial \hat{A}^{\mathrm{I}}(k,\ell)}, \tag{3.36}$$

$$\frac{\partial J^{\mathrm{I}}_{\mathrm{MSE}}\left(\hat{A}^{\mathrm{I}}(k,\ell),\hat{B}^{\mathrm{I}}(k,\ell)\right)}{\partial \hat{B}^{\mathrm{I}}(k,\ell)} = -2 \cdot Q^{\mathrm{I}}(k,\ell) \cdot \frac{\partial \hat{\Phi}^{\mathrm{I}}_{r_L}(k,\ell)}{\partial \hat{B}^{\mathrm{I}}(k,\ell)}. \tag{3.37}$$

The partial derivatives of the LRE PSD estimate $\hat{\Phi}^{\mathrm{I}}_{r_L}$ with respect to the parameter estimates will be computed for the OE and EE methods in subsections 3.6.1 and 3.6.2, respectively. It should be noted that when minimizing the MSE cost function, the parameter updates in each frame depend on the error Q^{I} between the LRE PSD and its estimate.

For both estimation methods, the gradient-descent update rule for the MSLE cost function is given in the logarithmic domain[1] as:

$$\boxed{\ln \underline{\hat{\theta}}^{\mathrm{I}}(k,\ell+1) = \ln \underline{\hat{\theta}}^{\mathrm{I}}(k,\ell) - \frac{\Gamma}{2} \odot \underline{\nabla}^{\mathrm{I}}_{\mathrm{MSLE}}(k,\ell),} \tag{3.38}$$

where the gradient of the MSLE cost function $\underline{\nabla}^{\mathrm{I}}_{\mathrm{MSLE}}$ is composed of the partial derivatives of the MSLE cost function with respect to the logarithm of the parameter estimates:

$$\underline{\nabla}^{\mathrm{I}}_{\mathrm{MSLE}}(k,\ell) = \left[\begin{array}{cc} \frac{\partial J^{\mathrm{I}}_{\mathrm{MSLE}}\left(\ln \hat{A}^{\mathrm{I}}(k,\ell), \ln \hat{B}^{\mathrm{I}}(k,\ell)\right)}{\partial \ln \hat{A}^{\mathrm{I}}(k,\ell)} & \frac{\partial J^{\mathrm{I}}_{\mathrm{MSLE}}\left(\ln \hat{A}^{\mathrm{I}}(k,\ell), \ln \hat{B}^{\mathrm{I}}(k,\ell)\right)}{\partial \ln \hat{B}^{\mathrm{I}}(k,\ell)} \end{array} \right]^T. \tag{3.39}$$

Using (3.29), these partial derivatives are equal to:

$$\frac{\partial J^{\mathrm{I}}_{\mathrm{MSLE}}\left(\ln \hat{A}^{\mathrm{I}}(k,\ell), \ln \hat{B}^{\mathrm{I}}(k,\ell)\right)}{\partial \ln \hat{A}^{\mathrm{I}}(k,\ell)} = -2 \cdot \left[\frac{Q^{\mathrm{I}}_{\ln}(k,\ell)}{\hat{\Phi}^{\mathrm{I}}_{r_L}(k,\ell)} \right] \cdot \frac{\partial \hat{\Phi}^{\mathrm{I}}_{r_L}(k,\ell)}{\partial \ln \hat{A}^{\mathrm{I}}(k,\ell)}, \tag{3.40}$$

$$\frac{\partial J^{\mathrm{I}}_{\mathrm{MSLE}}\left(\ln \hat{A}^{\mathrm{I}}(k,\ell), \ln \hat{B}^{\mathrm{I}}(k,\ell)\right)}{\partial \ln \hat{B}^{\mathrm{I}}(k,\ell)} = -2 \cdot \left[\frac{Q^{\mathrm{I}}_{\ln}(k,\ell)}{\hat{\Phi}^{\mathrm{I}}_{r_L}(k,\ell)} \right] \cdot \frac{\partial \hat{\Phi}^{\mathrm{I}}_{r_L}(k,\ell)}{\partial \ln \hat{B}^{\mathrm{I}}(k,\ell)}. \tag{3.41}$$

The partial derivatives of the LRE PSD estimate $\hat{\Phi}^{\mathrm{I}}_{r_L}$ with respect to the logarithm of the parameter estimates will be computed for the OE and EE methods in subsections 3.6.1 and 3.6.2, respectively. It should be noted that when minimizing the MSLE cost function, the parameter updates in each frame are normalized by the LRE PSD estimate $\hat{\Phi}^{\mathrm{I}}_{r_L}$ and depend on the log error $Q^{\mathrm{I}}_{\ln}$, which in turn depends on the ratio of the LRE PSD and its estimate.

3.6.1 *Algorithms for output error method*

Using (3.22), the partial derivatives of $\hat{\Phi}^{\mathrm{O}}_{r_L}$ with respect to the parameter estimates are equal to:

$$\begin{aligned} \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell)}{\partial \hat{A}^{\mathrm{O}}(k,\ell)} &= \Phi_x(k,\ell-G) + \hat{B}^{\mathrm{O}}(k,\ell) \cdot \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial \hat{A}^{\mathrm{O}}(k,\ell)}, \\ \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell)}{\partial \hat{B}^{\mathrm{O}}(k,\ell)} &= \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1) + \hat{B}^{\mathrm{O}}(k,\ell) \cdot \frac{\partial \hat{\Phi}^{O}_{r_L}(k,\ell-1)}{\partial \hat{B}^{O}(k,\ell)}, \end{aligned} \tag{3.42}$$

1 It should be noted that the gradient-descent update rule for the MSLE cost function in the linear domain yielded unreliable results.

while the partial derivatives of $\hat{\Phi}^{\mathrm{O}}_{r_L}$ with respect to the logarithm of the parameter estimates are equal to:

$$\begin{aligned}\frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell)}{\partial \ln \hat{A}^{\mathrm{O}}(k,\ell)} &= \hat{A}^{\mathrm{O}}(k,\ell)\cdot \Phi_x(k,\ell-G) + \hat{B}^{\mathrm{O}}(k,\ell)\cdot \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial \ln \hat{A}^{\mathrm{O}}(k,\ell)},\\ \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell)}{\partial \ln \hat{B}^{\mathrm{O}}(k,\ell)} &= \hat{B}^{\mathrm{O}}(k,\ell)\cdot \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1) + \hat{B}^{\mathrm{O}}(k,\ell)\cdot \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial \ln \hat{B}^{\mathrm{O}}(k,\ell)}.\end{aligned} \tag{3.43}$$

It should be noted that (3.42) and (3.43) contain partial derivatives of the OE PSD estimate $\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)$ in the *previous* frame with respect to the parameter estimates $\underline{\hat{\theta}}^{\mathrm{O}}(k,\ell)$ and their logarithm $\ln \underline{\hat{\theta}}^{\mathrm{O}}(k,\ell)$ in the *current* frame, respectively. These terms appear due to the recursive filter structure of the OE method. These partial derivatives cannot be computed in a straightforward manner, as $\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)$ does not directly depend on $\underline{\hat{\theta}}^{\mathrm{O}}(k,\ell)$. In [75], two approximations have been proposed for computing these partial derivatives, which we now apply to the problem at hand.

3.6.1.1 *Recursive prediction error (RPE)*

Although the OE PSD estimate $\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)$ in the previous frame does not directly depend on the parameter estimates $\underline{\hat{\theta}}^{\mathrm{O}}(k,\ell)$ in the current frame, it obviously directly depends on the parameter estimates $\underline{\hat{\theta}}^{\mathrm{O}}(k,\ell-1)$ in the previous frame. For computing the partial derivatives in (3.42) and (3.43), the RPE adaptive algorithm [75] uses the following approximations:

$$\begin{aligned}\frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial \underline{\hat{\theta}}^{\mathrm{O}}(k,\ell)} &\approx \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial \underline{\hat{\theta}}^{\mathrm{O}}(k,\ell-1)},\\ \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial \ln \underline{\hat{\theta}}^{\mathrm{O}}(k,\ell)} &\approx \frac{\partial \hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial \ln \underline{\hat{\theta}}^{\mathrm{O}}(k,\ell-1)},\end{aligned} \tag{3.44}$$

which have been shown to be reasonable if the step-sizes $\underline{\Gamma}$ in (3.34) and (3.38) are sufficiently small. Using these approximations makes it possible to compute the partial derivatives in (3.42) and (3.43) recursively. As a result, both reverberation parameters are updated even when the respective inputs to the parameters are absent.

3.6.1.2 *Pseudo linear regression (PLR)*

The PLR algorithm is an approximate gradient method [75] which assumes that the OE PSD estimate in the previous frame $\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)$ is independent of the parameter estimates in the current frame $\hat{\underline{\theta}}^{\mathrm{O}}(k,\ell)$, i.e.:

$$\begin{aligned}\frac{\partial\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial\hat{\underline{\theta}}^{\mathrm{O}}(k,\ell)}&=0,\\ \frac{\partial\hat{\Phi}^{\mathrm{O}}_{r_L}(k,\ell-1)}{\partial\ln\hat{\underline{\theta}}^{\mathrm{O}}(k,\ell)}&=0.\end{aligned}\tag{3.45}$$

Using (3.45) in (3.42) and (3.43) yields non-recursive formulations for the partial derivatives. It should be noted that the gradient computed using the PLR algorithm is an approximate version of the gradient computed using the RPE algorithm, as the assumptions in (3.45) are stronger than in (3.44).

3.6.2 *Algorithm for equation error method*

Using (3.30), the partial derivatives of $\hat{\Phi}^{\mathrm{E}}_{r_L}$ with respect to the parameter estimates are equal to:

$$\begin{aligned}\frac{\partial\hat{\Phi}^{\mathrm{E}}_{r_L}(k,\ell)}{\partial\hat{A}^{\mathrm{E}}(k,\ell)}&=\Phi_x(k,\ell-G),\\ \frac{\partial\hat{\Phi}^{\mathrm{E}}_{r_L}(k,\ell)}{\partial\hat{B}^{\mathrm{E}}(k,\ell)}&=\Phi_{r_L}(k,\ell-1),\end{aligned}\tag{3.46}$$

while the partial derivatives of $\hat{\Phi}^{\mathrm{E}}_{r_L}$ with respect to the logarithm of the parameter estimates are equal to:

$$\begin{aligned}\frac{\partial\hat{\Phi}^{\mathrm{E}}_{r_L}(k,\ell)}{\partial\ln\hat{A}^{\mathrm{E}}(k,\ell)}&=\hat{A}^{\mathrm{E}}(k,\ell)\cdot\Phi_x(k,\ell-G),\\ \frac{\partial\hat{\Phi}^{\mathrm{E}}_{r_L}(k,\ell)}{\partial\ln\hat{B}^{\mathrm{E}}(k,\ell)}&=\hat{B}^{\mathrm{E}}(k,\ell)\cdot\Phi_{r_L}(k,\ell-1).\end{aligned}\tag{3.47}$$

Hence, the partial derivatives obtained for the EE method in (3.46) and (3.47) are non-recursive and similar to those obtained for the PLR algorithm for the OE method. It can also be observed that the reverberation parameters are not updated when the respective inputs to the parameters are absent, i.e., the reverberation

scaling parameter $\hat{A}^{\mathrm{E}}$ is not updated when $\Phi_x(k, \ell - G) = 0$, while the reverberation decay parameter $\hat{B}^{\mathrm{E}}$ is not updated when $\Phi_{r_L}(k, \ell - 1) = 0$.

3.7 Simulations

In this section, we evaluate the performance of the proposed online parameter estimation methods (OE and EE), cost functions (MSE and MSLE) and gradient-descent-based algorithms, giving rise to 6 combinations: OE-RPE-MSE, OE-PLR-MSE, EE-MSE, OE-RPE-MSLE, OE-PLR-MSLE and EE-MSLE. In Sections 3.7.1 and 3.7.2 we describe the signals and the algorithmic parameters used in our simulations, while in Section 3.7.3 we discuss the performance metrics used to evaluate the PSD estimation accuracy, the residual echo suppression and the near-end speech distortion. In Section 3.7.4 we perform two experiments to evaluate the performance of the proposed parameter estimation methods. To evaluate the parameter estimation accuracy, the first experiment is performed in an idealistic setting using artificial RIRs with frequency-independent reverberation parameters. The second experiment is performed in a realistic setting using RIRs measured in different rooms, comparing the performance of the proposed methods with state-of-the-art signal-based methods.

3.7.1 *Signals*

In our simulations, we use time-domain signals at a sampling frequency $f_s = 16$ kHz. The far-end speech signal x of length 30s and the near-end speech signal s of length 5s are obtained from the TIMIT database [98], where the double-talk condition occurs in the last 5s. The background noise signal v of length 30s is stationary air conditioner noise measured in an office. The time-domain signals are transformed into the STFT domain with $N_{\mathrm{FFT}} = 512$ (i.e., $K = 257$) using a Hann analysis window and an overlap of 75%, i.e., a frameshift of $F = 128$.

The different RIRs used for our simulations can be divided into two categories:

- Artificial RIRs: A total of 30 RIRs were generated exactly according to the model in (3.12) with $N = 640$ and $N_h = 16000$ for all combinations of the frequency-independent parameters $\sigma_L^2 = \{-40, -36, -32, -28, -24, -20\}$ dB and $T_{60} = \{200, 400, 600, 800, 1000\}$ ms.
- Measured RIRs: A total of 55 RIRs were measured in 4 rooms with different reverberation times, with the number of RIRs measured in each room and the corresponding T_{60} values shown in Table 3.1. The broadband T_{60} of each RIR was estimated by line-fitting on its corresponding energy decay curve [29]. The lab, garage and the echoic room were rectangular shaped, while the office room was L-shaped. It should be noted that these RIRs obviously don't exactly correspond to the model in (3.12).

Room	**No. of IRs**	**T_{60}**
Lab	16	300-400 ms
Garage	16	400-500 ms
Office	16	500-600 ms
Echoic Room	7	850-950 ms

Table 3.1: Number of RIRs measured in each room and the corresponding reverberation times (T_{60}).

Method	MSLE		MSE	
	$\mu_{\mathbf{A}}$	$\mu_{\mathbf{B}}$	$\mu_{\mathbf{A}}$	$\mu_{\mathbf{B}}$
OE-RPE	10^{-2}	10^{-4}	10^{-4}	$10^{-3.5}$
OE-PLR	$10^{-1.75}$	$10^{-3.75}$	$10^{-2.5}$	10^{-2}
EE	10^{-1}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$

Table 3.2: Step-sizes used for the OE-RPE, OE-PLR and EE methods (for both the MSE and MSLE cost functions).

3.7.2 *Algorithmic parameters*

All required PSDs are computed via recursive smoothing according to (3.8), with the smoothing factor $\alpha = e^{\frac{-2 \cdot F}{f_s \cdot t_c}}$ computed for a time-constant $t_c = 0.02$ s. For the different combinations of parameter estimation methods, cost functions and gradient-descent-based algorithms, the step-sizes listed in Table 3.2 were used, which were found to give good results. In our experiments we however observed that the results obtained for the MSLE cost function were not very sensitive to the choice of the step-size. For the modified version of Favrot's method (see Appendix A.2), the delay M has been chosen as $M = N = 640$, while the delay P has been chosen as $P = \kappa \cdot F$ for two different values $\kappa = 12$ and $\kappa = 16$. In the RES postfilter in (3.9), an over-estimation factor $\beta = 2$ and a fixed spectral floor $\gamma = -20$ dB have been used.

3.7.3 *Performance metrics*

To evaluate the accuracy of the LRE PSD estimate $\hat{\Phi}_{r_L}$, we compute the Log Spectral Distance (LSD) [91] between the PSD estimate and the target PSD Φ_{r_L}, which can be expressed as the sum of the under- and over-estimation scores:

$$\mathrm{LSD} = \mathrm{LSD}_{\mathrm{un}} + \mathrm{LSD}_{\mathrm{ov}}, \tag{3.48}$$

$$\mathrm{LSD}_{\mathrm{un}} = \frac{10}{K \cdot L} \cdot \sum_{k=0}^{K-1} \sum_{\ell=l_1+1}^{l_1+L} \max\left\{0, \log_{10}\left(\frac{\Phi_{r_L}(k,\ell)}{\hat{\Phi}_{r_L}(k,\ell)}\right)\right\},$$

$$\mathrm{LSD}_{\mathrm{ov}} = -\frac{10}{K \cdot L} \cdot \sum_{k=0}^{K-1} \sum_{\ell=l_1+1}^{l_1+L} \min\left\{0, \log_{10}\left(\frac{\Phi_{r_L}(k,\ell)}{\hat{\Phi}_{r_L}(k,\ell)}\right)\right\},$$

where l_1 and L denote the start and the duration of the evaluation window (in frames), respectively. We choose l_1 corresponding to 20s ($l_1 = 2500$) and L corresponding to 5s ($L = 625$). A small LSD score corresponds to an accurate PSD estimate, with the perfect estimate $\hat{\Phi}_{r_L} = \Phi_{r_L}$ yielding LSD $= 0$.

To evaluate the amount of residual echo suppression and near-end speech distortion obtained by applying the RES postfilter, we compute the segmental residual echo attenuation (REA) and the segmental speech-to-speech distortion ratio (SSDR) [91, 99], respectively. The segmental REA is defined as:

$$\mathrm{REA}_{\mathrm{seg}} = \frac{1}{L} \cdot \sum_{\ell=l_1+1}^{l_1+L} \delta(\ell), \tag{3.49}$$

where

$$\delta(\ell) = 10 \cdot \log_{10}\left(\frac{\sum_{m=0}^{F-1} r_L^2(m+\ell \cdot F)}{\sum_{m=0}^{F-1} \tilde{r}_L^2(m+\ell \cdot F)}\right) \tag{3.50}$$

denotes the REA in each frame, with the late residual echo signal r_L and the processed residual echo signal $\tilde{r}_L$ obtained through inverse STFT processing of R_L and $\tilde{R}_L$ (see (3.11)), respectively. A large $\mathrm{REA}_{\mathrm{seg}}$ means that a large amount of residual echo has been suppressed. The segmental SSDR is defined as:

$$\mathrm{SSDR}_{\mathrm{seg}} = \frac{1}{L} \cdot \sum_{\ell=l_2+1}^{l_2+L} \eta(\ell), \tag{3.51}$$

where

$$\eta(\ell) = 10 \cdot \log_{10}\left(\frac{\sum_{m=0}^{F-1} s^2(m+\ell \cdot F)}{\sum_{m=0}^{F-1} \left(s(m+\ell \cdot F) - \tilde{s}(m+\ell \cdot F)\right)^2}\right) \tag{3.52}$$

denotes the SSDR in each frame, with s the near-end speech signal and $\tilde{s}$ the processed near-end speech signal (see (3.11)). Here, we choose l_2 corresponding to 25s ($l_2 = 3125$), such that the segmental SSDR is computed in the last 5s when double-talk occurs. A large $\mathrm{SSDR}_{\mathrm{seg}}$ corresponds to a small near-end speech signal distortion. In general, a trade-off exists between obtaining large residual echo attenuation and small near-end speech distortion. Hence, it is desirable to maximize $\mathrm{REA}_{\mathrm{seg}}$ while keeping $\mathrm{SSDR}_{\mathrm{seg}}$ as large as possible.

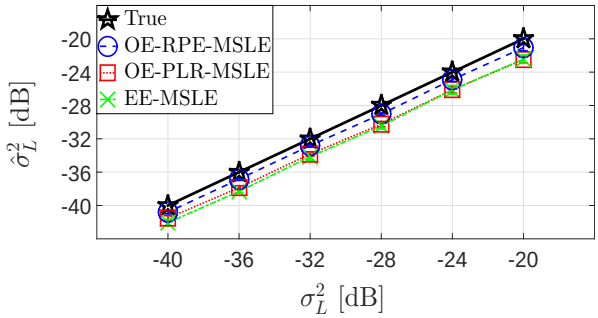

Fig. 3.6: Plot of $\hat{\sigma}_L^2$ vs σ_L^2 for the OE-RPE, OE-PLR and EE methods when minimizing the MSLE cost function for the idealistic setting.

3.7.4 *Experimental results*

The first experiment is performed in an idealistic setting, i.e., using artificial RIRs, a perfect AEC filter, no near-end speech and no background noise. In this experiment we evaluate how accurately the proposed methods estimate the RIR parameters and the LRE PSD. The second experiment is performed in a realistic setting using measured RIRs, a converged (but not perfect) subband AEC filter, near-end speech and background noise. In this experiment, we compare the LSD, segmental REA and SSDR scores and the T_{60} estimates obtained using the proposed online methods with those obtained using state-of-the-art methods, i.e., Valero's method [23] (offline version) and Favrot's method [24] (modified online version presented in Appendix A.2)

3.7.4.1 *Idealistic setting*

As already mentioned, in this experiment we use artificial RIRs with frequency-independent parameters σ_L^2 and T_{60} (see Section 3.7.1) to generate the acoustic echo signal and we assume a perfect AEC filter, i.e., no early residual echo ($R_E = 0$). Additionally, we assume that no near-end speech and background noise are present, i.e., $s(n) = v(n) = 0$, such that $E(k,\ell) = R_L(k,\ell)$. For this idealistic setting, we compare the estimates of the RIR parameters $\hat{\sigma}_L^2$ and $\hat{T}_{60}$ with the true values, and compare the LSD scores of the LRE PSD estimates obtained using the OE-RPE, OE-PLR and EE methods (for both the MSE and MSLE cost functions). For each method, the parameter estimates $\hat{\sigma}_L^2$ and $\hat{T}_{60}$ are obtained by averaging the converged values of the estimated model parameters $A(k)$ and $B(k)$ over all frequency bins and using them in (3.15), (3.16) and (3.13).

Fig. 3.6 and Fig. 3.7 show the estimated scaling parameter $\hat{\sigma}_L^2$ as a function of the true scaling parameter σ_L^2 for the OE-RPE, OE-PLR and EE methods when minimizing the MSLE and MSE cost functions, respectively. Each point in these figures corresponds to the average result obtained for 5 RIRs (with different T_{60}

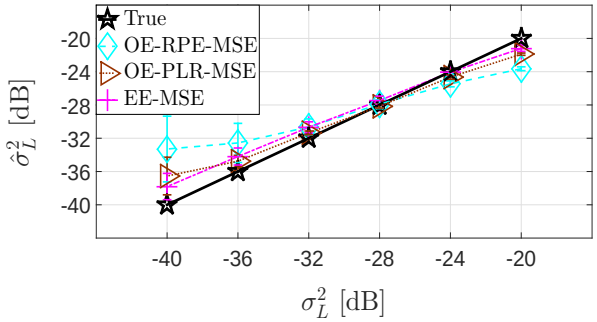

Fig. 3.7: Plot of $\hat{\sigma}_L^2$ vs σ_L^2 for the OE-RPE, OE-PLR and EE methods when minimizing the MSE cost function for the idealistic setting.

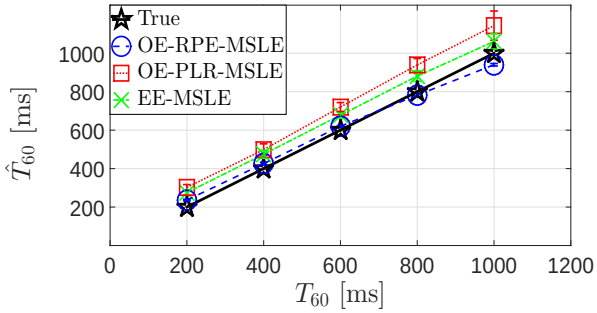

Fig. 3.8: Plot of $\hat{T}_{60}$ vs T_{60} for the OE-RPE, OE-PLR and EE methods when minimizing the MSLE cost function for the idealistic setting.

values), while the error bars depict the standard deviation across these 5 RIRs. On the one hand, it can be observed that for MSLE minimization (Fig. 3.6), all considered methods slightly under-estimate σ_L^2 and yield a very small standard deviation, indicating robustness to different T_{60} values. On the other hand, for MSE minimization (Fig. 3.7), all considered methods yield less accurate estimates with large standard deviations. Overall, the OE-RPE method with MSLE minimization gives the most accurate results for all considered σ_L^2 and T_{60}.

Fig. 3.8 and Fig. 3.9 show the estimated reverberation time $\hat{T}_{60}$ as a function of the true reverberation time T_{60} for the OE-RPE, OE-PLR and EE methods when minimizing the MSLE and MSE cost functions, respectively. Each point in these figures now corresponds to the average result obtained for 6 IRs (with different σ_L^2), while the error bars depict the standard deviation across these 6 IRs. It can be observed that for MSLE minimization (Fig. 3.8), the OE-RPE method estimates the T_{60} very accurately, while the OE-PLR and EE methods slightly over-estimate the T_{60}. All three methods yield small standard deviations, indicating robustness to different σ_L^2 values. For the MSE minimization (Fig. 3.9), the OE-RPE and OE-PLR methods estimate the T_{60} reasonably accurately with large standard deviations, while the EE method fails completely, especially for large T_{60}. Overall, the OE-RPE

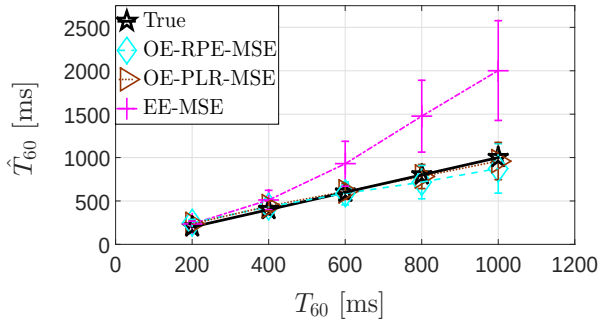

Fig. 3.9: Plot of $\hat{T}_{60}$ vs T_{60} for the OE-RPE, OE-PLR and EE methods when minimizing the MSE cost function for the idealistic setting.

Method	LSD_{un}					LSD_{ov}				
	$T_{60} = 0.2$s	0.4s	0.6s	0.8s	1s	0.2s	0.4s	0.6s	0.8s	1s
OE-RPE-MSLE	0.84	0.98	1.07	1.19	1.28	1.24	1.36	1.47	1.54	1.63
OE-PLR-MSLE	1.37	1.57	1.67	1.63	1.59	1.01	0.96	1.00	1.09	1.13
EE-MSLE	1.83	2.18	2.44	2.45	2.48	0.67	0.67	0.64	0.69	0.67
OE-RPE-MSE	1.15	0.90	0.95	0.99	1.02	1.25	1.97	2.51	2.93	3.10
OE-PLR-MSE	0.73	0.86	0.91	0.94	0.97	1.39	1.82	2.22	2.4	2.42
EE-MSE	0.81	0.41	0.18	0.08	0.06	1.38	2.9	5.27	7.66	9.01

Table 3.3: Average LSD scores obtained for artificially generated RIRs for all proposed parameter estimation methods.

method with MSLE minimization gives the most accurate and consistent results for all considered σ_L^2 and T_{60}.

Fig. 3.10 shows the LSD scores of the LRE PSD estimates obtained using all considered methods as a function of T_{60}. Each point in this figure again corresponds to the average result obtained for 6 RIRs (with different σ_L^2), while the error bars depict the standard deviation across these 6 RIRs. Additionally, Table 3.3 breaks down all average LSD scores into under- and over-estimation scores (see (3.48)). From these results it can be observed that the OE-RPE-MSLE and OE-PLR-MSLE methods consistently outperform all other methods across all T_{60} values, yielding the lowest LSD scores with the smallest standard deviations. When minimizing the MSE cost function, all methods yield significantly larger over-estimation scores than under-estimation scores, especially for large T_{60} values.

In conclusion, based on the results obtained for the idealistic setting, the OE-RPE-MSLE method outperforms all other proposed methods in terms of estimation accuracy of the RIR parameters σ_L^2 and T_{60} and the LRE PSD Φ_{r_L}. This corresponds to the result obtained in [77] for offline processing.

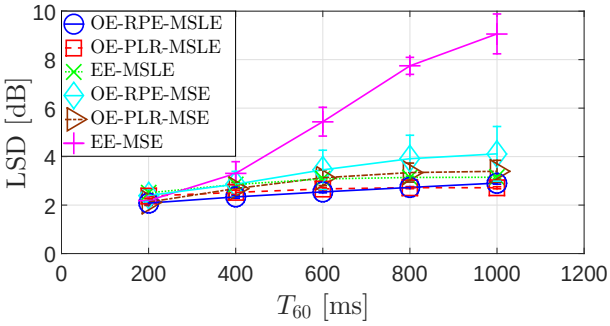

Fig. 3.10: Plot of LSD vs T_{60} for all proposed parameter estimation methods for the idealistic setting.

3.7.4.2 *Realistic setting*

In this experiment, we use measured RIRs (see Table 3.1) to generate the acoustic echo signal and subband AEC filter to perform echo cancellation (see Section 3.3.1). For the AEC filter we have used a rather short filter length ($G = 5$ frames, corresponding to 64 ms), aiming at canceling the direct sound component and the early reflections, while achieving fast convergence at low computational cost. The subband filter was pre-converged using the NLMS algorithm [15], with white Gaussian noise as the far-end signal. It should be noted that when using a subband AEC filter, the early residual echo is not completely cancelled, i.e., a small amount of early residual echo remains due to filter misalignment ($R_E \neq 0$). In addition, near-end speech and background noise are present, with the near-end signal-to-noise ratio set to 40 dB. In order to obtain a fair comparison of the segmental performance metrics for all measured RIRs, all RIRs have been scaled appropriately such that the speech-to-residual echo ratio (SRER) is equal to 10 dB. The reverberation parameters $\underline{\theta}(k)$ are estimated only during periods of near-end speech absence, i.e., during the first 25s, and when the AEC error PSD Φ_e is at least 3 dB above the background noise PSD Φ_v, as during these periods the AEC error PSD Φ_e is predominantly composed of the LRE PSD Φ_{r_L}. As Φ_{r_L} is not directly observable in practice, it is approximated in (3.30) by Φ_e during these periods.

For this realistic setting, we compare the LSD, REA_{seg} and SSDR_{seg} scores obtained using the OE-RPE, OE-PLR and EE methods (for both the MSE and MSLE cost functions) with the state-of-the-art methods in [23] (offline version) and [24] (modified online version). Additionally, we also compare the estimated reverberation time $\hat{T}_{60}$ with the (true) T_{60} obtained by line-fitting.

Fig. 3.11 shows the LSD scores obtained using all considered methods for the measured RIRs in each room. Each point in this figure corresponds to the average LSD score obtained for all RIRs in a specific room, while the error bars depict the standard deviation across these RIRs. It can be observed that the proposed OE-RPE-MSLE and OE-PLR-MSLE methods outperform all other online parameter

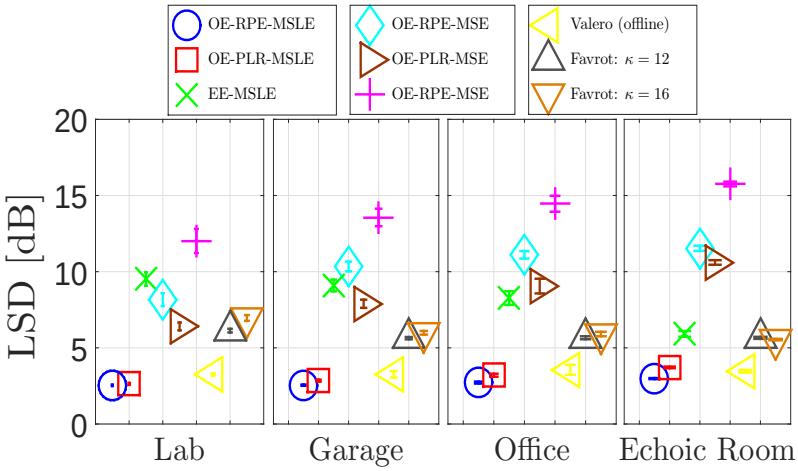

Fig. 3.11: LSD scores obtained for RIRs measured in 4 different rooms for all considered parameter estimation methods.

Method	Lab		Garage		Office		Echoic Room	
	LSD_{un}	LSD_{ov}	LSD_{un}	LSD_{ov}	LSD_{un}	LSD_{ov}	LSD_{un}	LSD_{ov}
OE-RPE-MSLE	1.00	1.54	1.03	1.52	1.17	1.58	1.40	1.63
OE-PLR-MSLE	1.30	1.33	1.26	1.59	1.90	1.30	1.45	2.27
EE-MSLE	0.23	9.29	0.20	8.89	0.28	7.99	0.49	5.43
OE-RPE-MSE	0.43	7.73	0.39	9.95	0.50	10.61	0.63	10.90
OE-PLR-MSE	0.58	5.83	0.67	7.22	0.80	8.26	0.86	9.74
EE-MSE	0.29	11.73	0.13	13.43	0.12	14.34	0.03	15.73
Valero (offline)	0.97	2.28	1.22	2.04	1.91	1.65	2.02	1.44
Favrot: $\kappa = 12$	0.77	5.36	0.72	4.91	0.92	4.75	1.28	4.37
Favrot: $\kappa = 16$	0.62	6.35	0.61	5.38	0.76	5.14	1.12	4.43

Table 3.4: Average LSD scores obtained for RIRs measured in 4 rooms (see Table 3.1) for all considered parameter estimation methods.

estimation methods, and are even slightly better than the offline method proposed in [23]. In addition, Table 3.4 breaks down all average LSD scores into under- and over-estimation scores. Firstly, it can be observed that among all proposed estimation methods, the OE-RPE-MSLE and OE-PLR-MSLE methods yield similar under- and over-estimation scores. Although the other proposed methods and the modified Favrot method yield smaller under-estimation scores than the OE-RPE-MSLE and OE-PLR-MSLE methods, they yield considerably larger over-estimation scores. Finally, for the offline method proposed in [23], both the under- and over-estimation scores are slightly larger than for the online OE-RPE-MSLE and OE-PLR-MSLE methods (except for under-estimation scores in the lab and over-estimation scores in the echoic room).

Fig. 3.12 shows the REA_{seg} scores against the SSDR_{seg} scores obtained using all considered methods. Each point in this figure corresponds to the average result

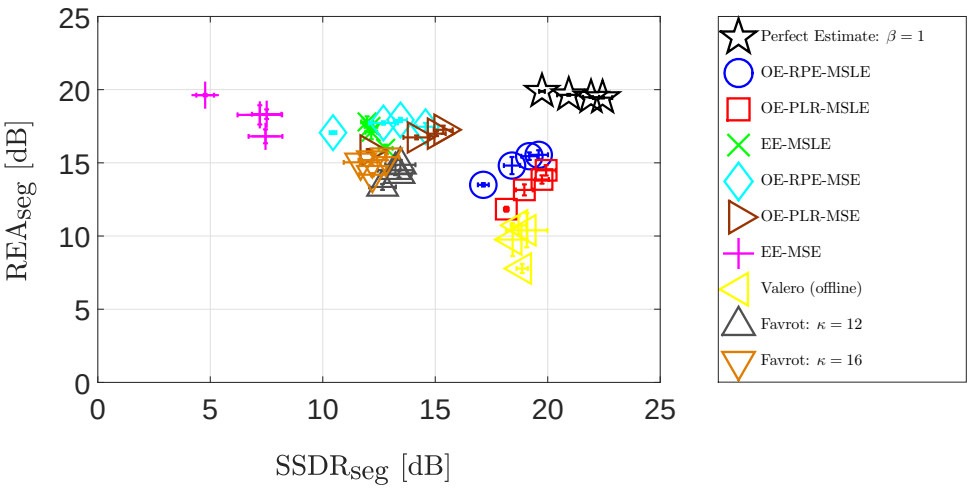

Fig. 3.12: Plot of segmental REA vs segmental SSDR obtained for RIRs measured in 4 different rooms for all considered parameter estimation methods.

obtained for all RIRs in a specific room, while the error bars on the x and y-axes depict the standard deviations across these RIRs for the SSDR$_{\text{seg}}$ and REA$_{\text{seg}}$ scores, respectively. For comparison, we also included the results obtained using the perfect LRE PSD estimate $\hat{\Phi}_{r_L} = \Phi_{r_L}$ and an over-estimation factor $\beta = 1$, which yields the best possible performance in terms of maximizing both the REA$_{\text{seg}}$ and SSDR$_{\text{seg}}$ scores. As expected, it can be observed that a large LSD over-estimation score (see Table 3.4) leads to large residual echo attenuation at the expense of large near-end speech distortion, while an opposite effect can be observed for a large LSD under-estimation score. The proposed online OE-RPE-MSLE and OE-PLR-MSLE methods as well as Valero's offline method yield significantly better SSDR$_{\text{seg}}$ scores as compared to the other methods (about 5-10 dB), while not losing too much in terms of the REA$_{\text{seg}}$ score (about 2-3 dB). Overall, the OE-RPE-MSLE method yields the best performance amongst all considered parameter estimation methods, i.e., both its REA$_{\text{seg}}$ as well as its SSDR$_{\text{seg}}$ score are closest to the scores obtained for the perfect LRE PSD estimate.

Fig. 3.13 shows the estimated reverberation time $\hat{T}_{60}$ obtained using all considered methods for the measured RIRs in each room. Each point in this figure corresponds to the average result obtained for all RIRs in a specific room, while the error bars depict the standard deviation across these RIRs. For comparison, the (true) T_{60} values obtained by line-fitting on the measured RIRs have also been included. It can be clearly observed that the OE-RPE-MSLE method yields the most accurate and consistent T_{60} estimate across all rooms. On the one hand, the OE-PLR-MSLE method, Valero's method and Favrot's method perform rather similarly, i.e., slightly over-estimating the T_{60} for the lower range (250-500 ms) but under-estimating the T_{60} for the higher range (600-900 ms). On the other hand, the EE method for both cost functions fails completely and significantly over-estimates the T_{60}, while the

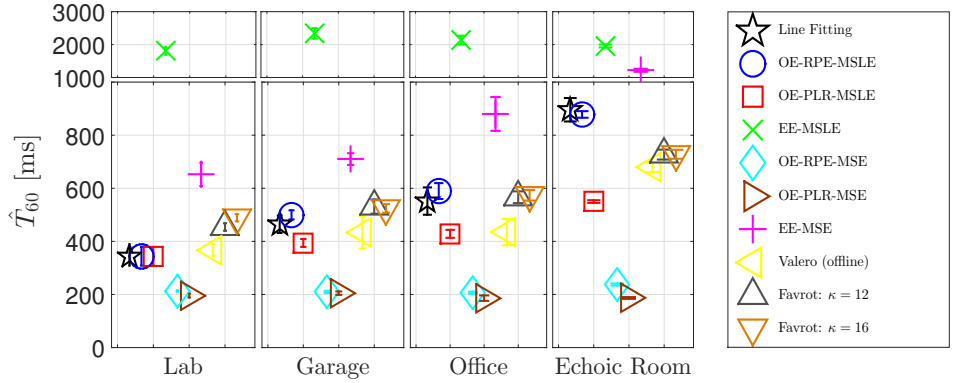

Fig. 3.13: Plot of $\hat{T}_{60}$ vs T_{60} (line-fitting) for RIRs measured in 4 different rooms for all considered parameter estimation methods.

OE-RPE-MSE and OE-PLR-MSE methods significantly underestimate the T_{60}. Additionally, in Fig. 3.14 we plot the estimated and true T_{60} values for all 55 measured RIRs and compute the correlation coefficient ζ between these values for each considered method. It can be seen that the proposed OE-RPE-MSLE method yields the largest correlation coefficient ($\zeta = 0.96$), followed by the proposed OE-PLR-MSLE method ($\zeta = 0.95$) and Favrot's method ($\zeta = 0.94$).

In conclusion, based on the results obtained for this realistic setting, the proposed OE-RPE-MSLE method outperforms all other considered (online and offline) parameter estimation methods in terms of LRE PSD and T_{60} estimation accuracy, while yielding the largest SSDR_{seg} score and hardly compromising on the REA_{seg} score compared to the perfect LRE PSD estimate.

3.8 Conclusion

In this paper, we considered late residual echo suppression by jointly estimating the parameters of an exponentially decaying reverberation model using online signal-based methods. The OE and EE methods, which were originally proposed to estimate the coefficients of time-domain IIR filters, were used on PSDs to jointly estimate the reverberation scaling and decay parameters by minimizing either the MSE or the MSLE cost function. For both methods, gradient-descent-based algorithms were derived to simultaneously update both parameters during periods of near-end speech absence. The estimated parameters were then used in a recursive filter structure to generate the corresponding LRE PSD estimate. The different methods (OE/EE), cost functions (MSE/MSLE) and gradient-descent-based algorithms (RPE/PLR) were compared with state-of-the-art signal-based methods, both in an idealistic as well as in a realistic setting. For both considered settings, the proposed OE-RPE-MSLE and OE-PLR-MSLE methods consistently outperformed all other considered methods in terms of LRE PSD estimation accuracy. Moreover, across all considered scenarios the OE-RPE-MSLE method yielded the most accurate T_{60} es-

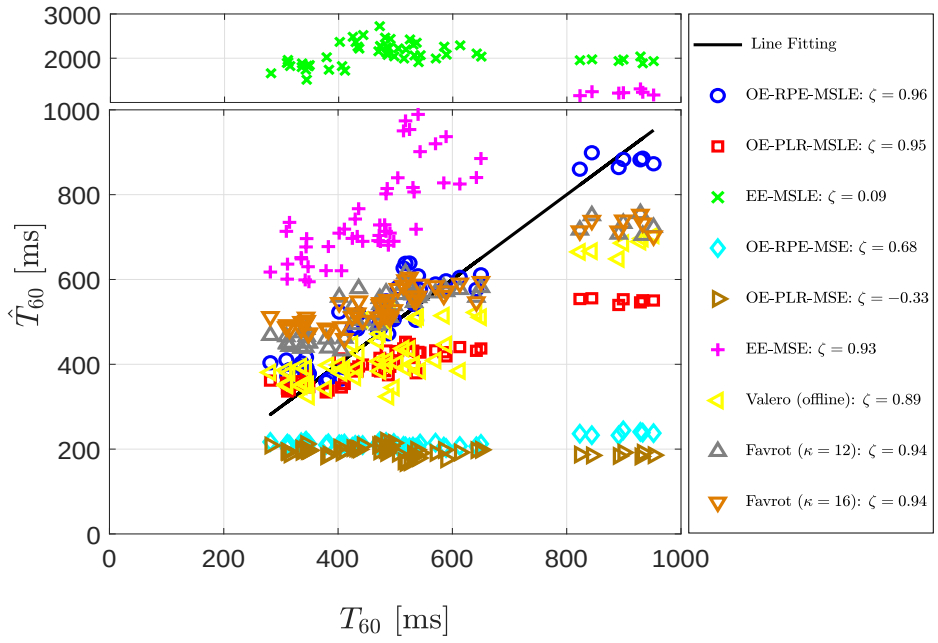

Fig. 3.14: Correlation between $\hat{T}_{60}$ obtained using each considered parameter estimation method and T_{60} (line-fitting) for all measured RIRs.

timates. The EE method failed to accurately estimate the LRE PSD and T_{60} across all scenarios, while both OE and EE methods for the MSE cost function failed to accurately estimate the T_{60}. For the realistic setting, the proposed OE-RPE-MSLE and OE-PLR-MSLE methods resulted in the smallest near-end speech distortion after applying the postfilter, while delivering a large residual echo suppression.

4

JOINT ONLINE ESTIMATION OF EARLY AND LATE RESIDUAL ECHO PSD FOR RESIDUAL ECHO SUPPRESSION

4.1 Abstract

In hands-free telephony and other distant-talking applications, an acoustic echo cancellation (AEC) system is typically required, where a short AEC filter is often used in practice to achieve fast convergence at low computational cost. This may result in late residual echo (LRE) remaining due to under-modeling of the echo path and early residual echo (ERE) due to filter misalignment. Both residual echo components can be suppressed using a postfilter in the subband domain, which requires accurate estimates of the power spectral density (PSD) of the ERE and LRE components. The ERE PSD has traditionally been estimated by proper scaling of the loudspeaker PSD, while a recursive estimator based on frequency-dependent reverberation scaling and decay parameters has frequently been used to estimate the LRE PSD. State-of-the-art methods estimate the required model parameters independently of each other. In this article, we propose to extend the ERE PSD estimator from a scalar to a moving average filter on the loudspeaker PSD, while the LRE PSD is estimated using an IIR filter based on the reverberation scaling and decay parameters. In addition, we propose a signal-based method to jointly estimate all model parameters in online mode, and derive two algorithms to simultaneously update the parameters by minimizing the mean squared log error. The proposed methods are compared with state-of-the-art methods in terms of estimation accuracy of the model parameters as well as the residual echo PSDs. Extensive simulation results using both artificially generated as well as measured impulse responses show that the proposed methods outperform state-of-the-art methods for all considered scenarios.

4.2 Introduction

Hands-free telephony applications and distant-talking applications, such as speech-enabled multimedia devices, have become very popular in recent years. In these

applications, the distance between the desired (near-end) speaker and the microphone may be quite large, while the loudspeaker playing back the far-end signal is typically located much closer to the microphone. As a result, the microphone signal may be degraded significantly due to the acoustic echo of the far-end signal, which may lead to the near-end speaker being unintelligible. In a typical acoustic echo cancellation (AEC) system, an adaptive filter aims at estimating the impulse response (IR) between the loudspeaker and the microphone [1, 10, 11]. In practice, however, the filter is typically not able to perfectly estimate the IR, resulting in residual echo due to filter misalignment. Additionally, as a short filter is often used in practice in order to achieve fast convergence at low computational cost, the filter is unable to estimate the complete echo path, leading to late residual echo. Thus, assuming no non-linear signal components, the residual echo is composed of early residual echo (ERE) due to filter misalignment and late residual echo (LRE) due to under-modeling of the IR by the short AEC filter.

The residual echo is often suppressed in the subband domain using a postfilter [21, 22, 62–64, 66, 67], which relies on an estimate of the power spectral density (PSD) of the residual echo. Hence, it is desirable to accurately estimate the PSD of both the ERE and LRE components. A simple but frequently used method is to estimate the ERE PSD as a scaled version of the PSD of either the far-end signal [1] or the estimated echo signal (generated by the AEC filter) [65]. In either case, the scalar is estimated during periods of near-end speech absence by computing a ratio between the PSD of the AEC error signal (obtained after the estimated echo signal is subtracted from the microphone signal) and the PSD of the respective input signal.

To estimate the LRE PSD, several methods have been proposed based on the statistical reverberation model in [28], which assumes that the late reverberant part of an IR decays exponentially at a rate proportional to the reverberation time. A recursive estimator for the LRE PSD was proposed in [13], which requires estimates of two frequency-independent room acoustic parameters: the reverberation scaling parameter (which is related to the initial power of the LRE component) and the reverberation decay parameter (which is related to the reverberation time). Both reverberation parameters were estimated using a *channel-based* method, i.e. using the coefficients of the converged AEC filter. In [14], a similar recursive estimator for the LRE PSD was derived with frequency-dependent reverberation parameters, where both parameters were again estimated using a channel-based method. It should be noted that channel-based methods are effective only if the AEC filter is long enough to capture a significant portion of the decay of the IR. Hence, *signal-based* methods have also been proposed, which estimate the reverberation parameters using the far-end and residual echo signals. A recursive estimator for the LRE PSD was derived in [23] based on the generalized reverberation model in [72], where a signal-based method was proposed to estimate the reverberation parameters in offline mode (i.e. batch processing). In [77] and [78], we proposed two signal-based methods to *jointly* estimate both reverberation parameters (in offline and online mode) by minimizing either the mean squared error (MSE) or the mean squared log error (MSLE) cost function. In [24], a coupling-factor-based estimator for the early acoustic echo

PSD and a recursive estimator for the late acoustic echo PSD were considered in a pure acoustic echo suppression system (i.e. without AEC filter). A signal-based method exploiting higher-order-statistics was proposed to estimate the parameters independently of each other in online mode.

As an extension of [78], in this paper, we propose signal-based methods to estimate the PSD of both residual echo components based on parametric models. By assuming that the filter misalignment is spread evenly over all AEC filter taps [1, 70, 71], we propose to model the ERE PSD using a moving average filter (instead of a scalar) on the PSD of the far-end signal, based on a frequency-dependent coupling factor. Similarly as in [14, 78], the LRE PSD is modeled using an IIR filter on the PSD of the far-end signal based on (frequency-dependent) reverberation scaling and decay parameters. We propose to jointly estimate all three model parameters (both reverberation parameters and the coupling factor) in online mode using the output error method by minimizing a single MSLE cost function. To simultaneously update the model parameters, we use gradient-descent-based algorithms such as recursive prediction error and pseudo-linear regression, which were originally derived for time-domain recursive systems [15, 75, 76]. The proposed methods are first evaluated in an idealistic setting, i.e. using artificially generated IRs and no AEC filter. They are then compared with state-of-the-art methods [1], [23] and [24] in a realistic setting, i.e. using IRs measured in different rooms and pre-converged AEC filters, in terms of estimation accuracy of the residual echo PSD and the resulting residual echo suppression and near-end speech distortion.

The remainder of the paper is organized as follows. The signal model as well as the AEC and postfilter systems are introduced in Section 4.3. In Section 4.4, the considered models for the ERE and LRE PSDs are presented. In Section 4.5.1, we discuss state-of-the-art methods for estimating the different model parameters. In Section 4.5.2, we present the proposed method for jointly estimating the model parameters by minimizing the MSLE cost function in online mode, with either of two gradient-descent-based algorithms used to simultaneously update the parameters. In Section 4.6, simulation results using artificial as well as measured IRs are presented.

4.3 Signal model, AEC and postfilter systems

Fig. 4.1 shows a loudspeaker-enclosure-microphone (LEM) system in which the far-end signal x is played through the loudspeaker and the microphone captures the acoustic echo component d, the near-end speech signal s and the background noise signal v. The microphone signal at discrete-time sample n is thus given as:

$$y(n) = s(n) + v(n) + \underbrace{\sum_{i=0}^{N_h-1} h(i) \cdot x(n-i)}_{d(n)}, \tag{4.1}$$

where h denotes the IR between the loudspeaker and the microphone, which is assumed to be time-invariant and of length N_h samples. To remove the acoustic

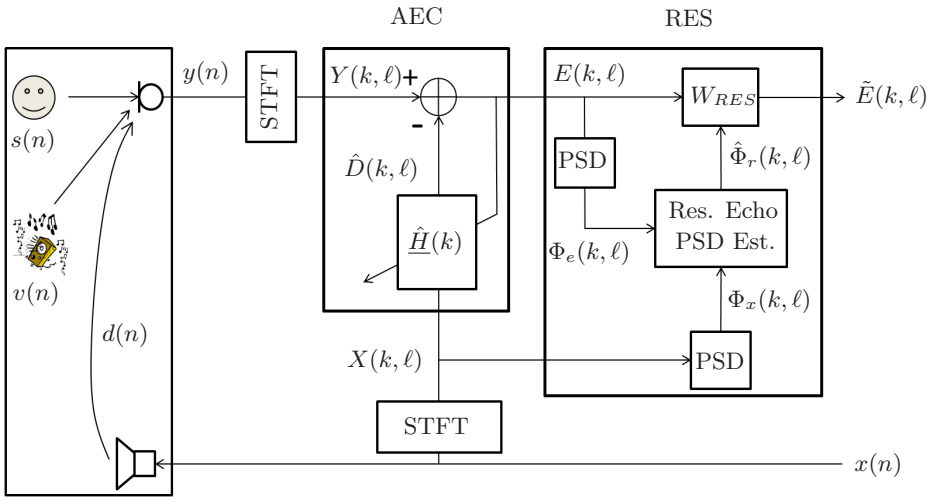

Fig. 4.1: Acoustic echo cancellation (AEC) and residual echo suppression (RES) systems.

echo component from the microphone signal, we consider a system in the subband domain consisting of two parts: an AEC filter $\underline{\hat{H}}$ and a residual echo suppression (RES) filter W_{RES}.

4.3.1 *Acoustic echo cancellation*

We consider a G-tap subband AEC filter $\underline{\hat{H}}$, with the filter length G chosen so as to cover only the direct path and early reflections in h. For subband processing, the (windowed) time-domain signals are transformed into the short-time Fourier transform (STFT) domain using a fast Fourier transform (FFT) filterbank of order N_{FFT}, with the total number of subbands $K = \frac{N_{\mathrm{FFT}}}{2} + 1$. The complex-valued spectrum of the far-end signal x in the k^{th} subband and ℓ^{th} frame is given as:

$$X(k,\ell) = \sum_{i=0}^{N_{\mathrm{FFT}}-1} x(\ell \cdot F + i) \cdot W_{\mathrm{ana}}(i) \cdot e^{-j\frac{2\pi}{N_{\mathrm{FFT}}}ki}, \tag{4.2}$$

where $j = \sqrt{-1}$, F denotes the frameshift and W_{ana} denotes the analysis window. The spectra of the other time-domain signals are computed similarly to (4.2), with the spectral equivalent of (4.1) given as:

$$Y(k,\ell) = S(k,\ell) + V(k,\ell) + D(k,\ell). \tag{4.3}$$

The acoustic echo estimate is generated by filtering the far-end signal through the AEC filter:

$$\hat{D}(k,\ell) = \underline{X}^H(k,\ell)\ \underline{\hat{H}}(k), \tag{4.4}$$

where $\underline{X}(k,\ell) = \begin{bmatrix} X(k,\ell) & \dots & X(k,\ell-G+1) \end{bmatrix}^T$ denotes the G-dimensional input vector to the subband AEC filter $\underline{\hat{H}}$, $\cdot^H$ denotes the Hermitian operator and $\cdot^T$ denotes the transpose operator. The AEC error signal is then given as:

$$\begin{aligned} E(k,\ell) &= Y(k,\ell) - \hat{D}(k,\ell) \\ &= S(k,\ell) + V(k,\ell) + \Big(D(k,\ell) - \hat{D}(k,\ell)\Big) \\ &= S(k,\ell) + V(k,\ell) + R(k,\ell) \\ &= S(k,\ell) + V(k,\ell) + \underbrace{R_E(k,\ell)}_{\text{Misalignment}} + \underbrace{R_L(k,\ell)}_{\text{Under-modeling}}, \end{aligned} \tag{4.5}$$

where R, R_E and R_L denote the residual echo, ERE and LRE components, respectively. The ERE component is given as:

$$R_E(k,\ell) = \underline{X}^H(k,\ell)\ \Delta\underline{H}_E(k), \tag{4.6}$$

with the AEC misalignment filter defined as:

$$\Delta\underline{H}_E(k) = \underline{H}_E(k) - \underline{\hat{H}}(k), \tag{4.7}$$

where $\underline{H}_E$ contains the first G coefficients of the equivalent subband filter corresponding to h. Since in this paper $G = \lfloor \frac{N}{F} \rfloor$, where $N \ll N_h$ corresponds to the length of the direct path and early reflections in h, the LRE component R_L is assumed to contain only late reflections, also known as reverberation.

4.3.2 *Residual echo suppression*

From (4.5), it can be observed that in addition to the desired near-end speech signal, the AEC error signal also contains the background noise and residual echo components. It is desirable to suppress these interfering signals while maintaining high quality and low distortion of the near-end speech signal. As shown in Fig. 4.1, this suppression is performed by applying a real-valued postfilter W_{RES} to the AEC error signal E. A frequently used postfilter is the Wiener gain [1], i.e.:

$$W_{\text{RES}}(k,\ell) = 1 - \left(\frac{\lambda_r(k,\ell) + \lambda_v(k,\ell)}{\lambda_e(k,\ell)}\right), \tag{4.8}$$

where λ_r, λ_v and λ_e denote the PSDs of the residual echo, background noise and AEC error signals, respectively. Assuming that S, V and R are mutually uncorrelated, the PSD of the AEC error signal can be expressed using (4.5) as:

$$\lambda_e(k,\ell) = \mathcal{E}\left\{|E(k,\ell)|^2\right\} = \lambda_s(k,\ell) + \lambda_v(k,\ell) + \lambda_r(k,\ell), \tag{4.9}$$

where $\mathcal{E}\{\cdot\}$ denotes the statistical expectation operator. Additionally, we make the realistic assumption that the early and late reflections in the IR h are uncorrelated, such that the residual echo PSD can be written as:

$$\lambda_r(k,\ell) = \lambda_{r_E}(k,\ell) + \lambda_{r_L}(k,\ell), \tag{4.10}$$

where λ_{r_E} and λ_{r_L} denote the ERE PSD and LRE PSD, respectively.

In practice, the statistical expectation operator in (4.9) is approximated by temporal averaging, e.g.:

$$\Phi_e(k,\ell) = \alpha \cdot \Phi_e(k,\ell-1) + (1-\alpha) \cdot |E(k,\ell)|^2, \tag{4.11}$$

where Φ_e is an approximation of λ_e and α denotes a smoothing factor. For an unobservable signal such as R, the quantity Φ_r itself needs to be estimated, with its estimate denoted as $\hat{\Phi}_r$. In the remainder of this paper, we will use the term *true PSD* to refer to the quantity λ, the term *PSD* to refer to the quantity Φ and the term *PSD estimate* to refer to its estimate $\hat{\Phi}$.

In order to control the aggressiveness of the residual echo suppression, we will use the following gain:

$$\boxed{W_{\mathrm{RES}}(k,\ell) = \max\left\{1 - \beta \cdot \left(\frac{\hat{\Phi}_r(k,\ell) + \hat{\Phi}_v(k,\ell)}{\Phi_e(k,\ell)}\right),\ \gamma\right\},} \tag{4.12}$$

where β denotes the over-estimation factor and γ denotes the (fixed) spectral floor, i.e. the maximum attenuation of the filter. Many approaches have been proposed in literature to estimate the PSD of the background noise $\hat{\Phi}_v$ [89–91]. In this paper, we assume that the background noise is stationary and its PSD estimate $\hat{\Phi}_v$ is known. Based on (4.10), the residual echo PSD estimate is given by:

$$\boxed{\hat{\Phi}_r(k,\ell) = \hat{\Phi}_{r_E}(k,\ell) + \hat{\Phi}_{r_L}(k,\ell).} \tag{4.13}$$

Although during near-end speech absence $\hat{\Phi}_r$ can be easily estimated from Φ_e based on (4.9), this is obviously not possible during periods of double-talk. Hence, in this paper we will use parametric models for the ERE PSD Φ_{r_E} and the LRE PSD Φ_{r_L}, which will be explained in the next section.

The processed AEC error signal is given as:

$$\tilde{E}(k,\ell) = W_{\mathrm{RES}}(k,\ell) \cdot E(k,\ell), \tag{4.14}$$

which can be expressed as the sum of its individual components similarly to (4.5):

$$\tilde{E}(k,\ell) = \tilde{S}(k,\ell) + \tilde{V}(k,\ell) + \tilde{R}(k,\ell), \tag{4.15}$$

where $\tilde{S}$, $\tilde{V}$ and $\tilde{R}$ are obtained by independently filtering S, V and R respectively with W_{RES}. The processed signals $\tilde{E}$, $\tilde{S}$ and $\tilde{R}$ are synthesized into the time-domain using inverse STFT and overlap-add processing to yield the processed time-domain signals $\tilde{e}$, $\tilde{s}$ and $\tilde{r}$, respectively. These signals can then be used to compute metrics to evaluate the near-end speech distortion and residual echo suppression (see Section 4.6.3).

4.4 Models for early and late residual echo PSD

In this section, we present the considered models for the early and late residual echo PSDs. We propose to model the ERE PSD using a moving average filter on the PSD of the far-end signal. Similarly as in [14, 78], the LRE PSD is modeled using an IIR filter on the PSD of the far-end signal.

4.4.1 *Model for early residual echo PSD*

As already mentioned, the ERE is caused by the misalignment between the IR and the AEC filter. A simple model for the ERE PSD was proposed in [1], where the ERE PSD is modeled as a scaled version of the PSD of the far-end signal:

$$\hat{\Phi}_{r_E}(k,\ell) = C(k) \cdot \Phi_x(k,\ell), \tag{4.16}$$

where C denotes the (frequency-dependent) coupling factor. As shown in [1], the coupling factor represents the squared magnitude spectrum of the filter misalignment. A disadvantage of this model is that a scalar coupling factor may not be sufficient to model the ERE PSD, especially if a long AEC filter is used.

We now derive our proposed model for the ERE PSD. Using (4.6), the ERE PSD is given by:

$$\begin{aligned}
\lambda_{r_E}(k,\ell) &= \mathcal{E}\left\{|R_E(k,\ell)|^2\right\} \\
&= \mathcal{E}\left\{\left|\sum_{g=0}^{G-1} X^*(k,\ell-g)\cdot\Delta H_E(k,g)\right|^2\right\}, \\
&= \mathcal{E}\left\{\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} X^*(k,\ell-i)\cdot X(k,\ell-j)\cdot\Delta H_E(k,i)\cdot\Delta H_E^*(k,j)\right\},
\end{aligned} \tag{4.17}$$

where $\Delta H_E(k,g)$ denotes the g^{th} coefficient of the AEC misalignment filter $\Delta\underline{H}_E$. Assuming statistical independence between the far-end signal and the AEC misalignment filter yields:

$$\lambda_{r_E}(k,\ell) = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \mathcal{E}\left\{X^*(k,\ell-i)\cdot X(k,\ell-j)\right\}\cdot\mathcal{E}\left\{\Delta H_E(k,i)\cdot\Delta H_E^*(k,j)\right\}. \tag{4.18}$$

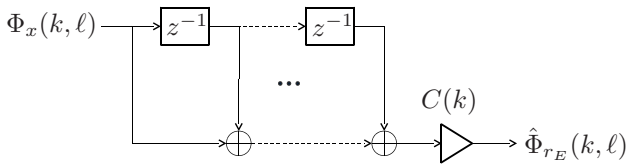

Fig. 4.2: Proposed model for the ERE PSD Φ_{r_E} (moving average filter).

Assuming that the coefficients of the AEC misalignment filter are mutually uncorrelated, i.e., $\mathcal{E}\{\Delta H_E(k,i) \cdot \Delta H_E^*(k,j)\} = 0$ for $i \neq j$, the ERE PSD can be written as:

$$\lambda_{r_E}(k,\ell) = \sum_{g=0}^{G-1} \lambda_x(k,\ell-g) \cdot \mathcal{E}\left\{|\Delta H_E(k,g)|^2\right\}. \tag{4.19}$$

Finally assuming that the misalignment is spread evenly over all AEC filter coefficients [1, 70, 71], i.e., $\mathcal{E}\left\{|\Delta H_E(k,g)|^2\right\} = C(k)\ \forall\ g$, the ERE PSD can be simplified as:

$$\lambda_{r_E}(k,\ell) = C(k) \cdot \sum_{g=0}^{G-1} \lambda_x(k,\ell-g). \tag{4.20}$$

Based on (4.20), we will hence use the following model for the ERE PSD:

$$\boxed{\hat{\Phi}_{r_E}(k,\ell) = C(k) \cdot \sum_{g=0}^{G-1} \Phi_x(k,\ell-g).} \tag{4.21}$$

This model can be interpreted as an extension of (4.16) in that a moving average filter is used instead of an instantaneous scaling of the PSD of the far-end signal. This model is depicted in Fig. 4.2.

4.4.2 *Model for late residual echo PSD*

As already mentioned, the LRE component is caused by under-modeling of the IR by the AEC filter. Several models for the LRE PSD have been proposed based on the statistical reverberation model in [28], which assumes that the late reverberant part of an IR can be described as an exponentially decaying realization of a stochastic process:

$$h(i) = w_L(i) \cdot e^{-\rho\ (i-N)}, \qquad N \leq i < N_h, \tag{4.22}$$

where $w_L \sim \mathcal{N}(0, \sigma_L^2)$ is a zero-mean stationary white Gaussian noise process with variance σ_L^2 and ρ denotes the decay rate. The decay rate is related to the reverberation time T_{60} of the room as:

$$\rho = \frac{3 \cdot \ln 10}{f_s \cdot T_{60}}, \tag{4.23}$$

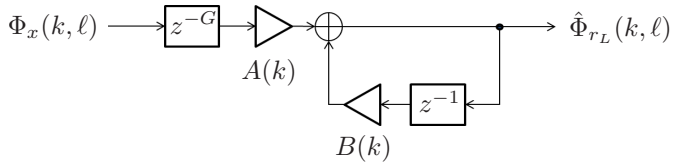

Fig. 4.3: Model for the LRE PSD Φ_{r_L} (IIR filter).

where f_s denotes the sampling rate in Hz. It should be noted that in practice the T_{60}, and hence the decay rate ρ, are frequency-dependent [25].

Based on (4.22), a recursive expression for the LRE PSD was first derived in [13] using frequency-independent parameters. In this paper, we will use a version of this model with frequency-dependent parameters, which was derived in [14] and [78], and is given as:

$$\boxed{\hat{\Phi}_{r_L}(k,\ell) = A(k)\cdot \Phi_x(k,\ell-G) + B(k)\cdot \hat{\Phi}_{r_L}(k,\ell-1),} \tag{4.24}$$

where $A(k)$ and $B(k)$ denote the frequency-dependent reverberation scaling and decay parameters, respectively. These parameters are related to the frequency-dependent variance $\sigma_L^2(k)$ and decay rate $\rho(k)$ as follows (see [78]):

$$A(k) = \sigma_L^2(k)\cdot\left(\frac{1-e^{-2\rho(k)F}}{1-e^{-2\rho(k)}}\right), \tag{4.25}$$

$$B(k) = e^{-2\rho(k)F}. \tag{4.26}$$

The recursive expression in (4.24) is depicted in Fig. 4.3 as an IIR filter on the PSD of the far-end signal.

4.5 Parameter estimation methods

In Section 4.5.1, we briefly review state-of-the-art signal-based methods to estimate the model parameters A, B and C. In Section 4.5.2, we present our proposed signal-based methods to jointly estimate all model parameters by minimizing a single cost function. Please note that in all the considered methods, the parameters are estimated only during periods of near-end speech absence.

4.5.1 *State-of-the-art methods*

In this section, we briefly discuss state-of-the-art methods for estimating the three model parameters A, B and C.

Hänsler et al. [1] estimate the scalar coupling factor C in (4.16) as the smoothed ratio of the AEC error PSD and the far-end signal PSD:

$$\hat{C}_{\mathrm{H}}(k,\ell) = (1-\delta)\cdot\frac{\Phi_e(k,\ell)}{\Phi_x(k,\ell)} + \delta\cdot\hat{C}_{\mathrm{H}}(k,\ell-1), \tag{4.27}$$

where δ denotes a smoothing factor. Please note that in [1], no additional estimator for the LRE PSD was used, i.e., the estimated coupling factor from (4.27) was fed into (4.16) to yield an estimate for the complete residual echo PSD $\hat{\Phi}_r$.

Valero et al. [23] proposed a method to estimate the reverberation parameters A and B by minimizing two different cost functions in offline (i.e., batch processing) mode. The reverberation decay parameter B was estimated by minimizing the MSLE cost function:

$$J_{\mathrm{MSLE}}\ \left(\ln\hat{B}_{\mathrm{V}}(k)\right) = \sum_{\ell=0}^{N_T-1}\left(\ln\Phi_e(k,\ell) - \ln\hat{\Phi}_{r_L}(k,\ell)\right)^2, \tag{4.28}$$

where N_T is the batch size in frames. The reverberation scaling parameter A was then estimated by minimizing the MSE cost function:

$$J_{\mathrm{MSE}}\ \left(\hat{A}_{\mathrm{V}}(k)\right) = \sum_{\ell=0}^{N_T-1}\left(\Phi_e(k,\ell) - \hat{\Phi}_{r_L}(k,\ell)\right)^2. \tag{4.29}$$

Please note that in [23], no additional estimator for the ERE PSD was used, i.e., the estimated reverberation parameters $\hat{A}_{\mathrm{V}}$ and $\hat{B}_{\mathrm{V}}$ were fed into (4.24) to yield an estimate for the complete residual echo PSD $\hat{\Phi}_r$.

Favrot et.al. [24] considered a pure acoustic echo suppression setup (i.e., no AEC filter) and proposed a coupling-factor-based estimator for the early acoustic echo PSD as well as a recursive estimator for the late acoustic echo PSD. The model parameters were estimated independently of each other in online mode using a method based on higher-order statistics. In order to facilitate a fair comparison, in this paper we consider a modified version of Favrot's method to estimate all three model parameters, and therefore both the ERE and LRE PSDs, in the presence of an AEC filter (see Appendix B.1).

4.5.2 *Joint parameter estimation methods*

Based on the parametric models for the ERE and LRE PSDs discussed in Section 4.4, in this paper we propose methods to *jointly* estimate all three model parameters A, B and C in *online* mode. These methods are extensions of the methods proposed in [78], which assumed no filter misalignment, i.e., $\Phi_{r_E} = 0$, and therefore only estimated the reverberation parameters A and B. To jointly estimate the parameters of generic IIR filters in the time-domain, several signal-based methods have been proposed [15,75,76,92–94], either based on output error (OE) or equation error (EE).

In [78] we investigated both the OE and EE methods (applied to PSDs) to jointly estimate the reverberation parameters A and B, either using the MSE or MSLE cost function. Simulation results showed that the OE method using the MSLE cost function yielded the best performance in terms of PSD estimation accuracy and residual echo suppression. Therefore, in this paper we will only consider the OE method using the MSLE cost function to jointly estimate all model parameters (reverberation parameters A and B and coupling factor C).

By merging the moving average model for the ERE PSD in (4.21) with the recursive model for the LRE PSD in (4.24), the residual echo PSD estimate is given as:

$$\begin{aligned}\hat{\Phi}_r(k,\ell) &= \hat{\Phi}_{r_E}(k,\ell) + \hat{\Phi}_{r_L}(k,\ell) \\ &= \hat{C}(k,\ell) \cdot \sum_{g=0}^{G-1} \Phi_x(k,\ell-g) + \\ &\quad \hat{A}(k,\ell) \cdot \Phi_x(k,\ell-G) + \hat{B}(k,\ell) \cdot \hat{\Phi}_{r_L}(k,\ell-1),\end{aligned} \tag{4.30}$$

where $\hat{A}(k,\ell)$, $\hat{B}(k,\ell)$ and $\hat{C}(k,\ell)$ denote estimates of the model parameters in subband k and frame ℓ and can be represented by the vector:

$$\underline{\hat{\theta}}(k,\ell) = \begin{bmatrix} \hat{A}(k,\ell) & \hat{B}(k,\ell) & \hat{C}(k,\ell) \end{bmatrix}^T. \tag{4.31}$$

From (4.30), it can be observed that the PSD estimate in the current frame $\hat{\Phi}_r(k,\ell)$ not only depends on the parameter estimates in the current frame $\underline{\hat{\theta}}(k,\ell)$ but also on the PSD estimate $\hat{\Phi}_r(k,\ell-1)$, which itself depends on the parameter estimates in the previous frame $\underline{\hat{\theta}}(k,\ell-1)$, and so on. Thus, $\hat{\Phi}_r$ is a non-linear function of $\underline{\hat{\theta}}$, where the current PSD estimate depends on parameter estimates in all previous frames.

The logarithmic error between the target PSD Φ_r and the PSD estimate $\hat{\Phi}_r$ in (4.30) is defined as:

$$Q_{\ln}(k,\ell) = \ln\left(\frac{\Phi_r(k,\ell)}{\hat{\Phi}_r(k,\ell)}\right). \tag{4.32}$$

To update all model parameters in each frame, we will consider the instantaneous MSLE cost function:

$$J\left(\ln\hat{A}(k,\ell),\ \ln\hat{B}(k,\ell),\ \ln\hat{C}(k,\ell)\right) = Q_{\ln}^2(k,\ell). \tag{4.33}$$

Similarly as in [78], we now derive gradient-descent-based algorithms to update the model parameters $\underline{\hat{\theta}}(k,\ell)$. Since the residual echo PSD Φ_r is obviously not observable, we will only update the model parameters during periods of near-end speech absence and when the AEC error signal is not dominated by background noise, such that we can replace Φ_r by Φ_e in (4.32). The parameters will then be used, both during periods of near-end speech absence as well as double-talk, to estimate the residual

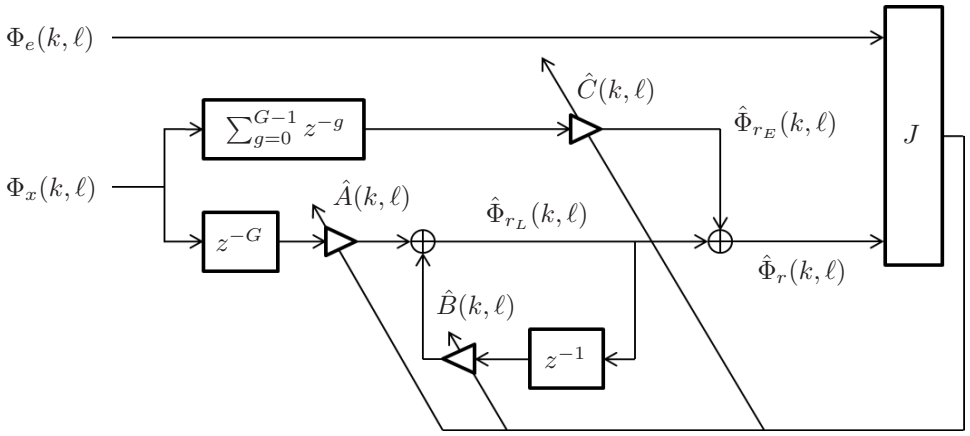

Fig. 4.4: Online joint estimation of the three model parameters using the output error method by minimizing a single MSLE cost function.

echo PSD $\hat{\Phi}_r$. The block scheme to estimate the model parameters in online mode is depicted in Fig. 4.4.

The gradient-descent update rule for J in the logarithmic domain is given as:

$$\boxed{\ln\hat{\theta}(k,\ell+1) = \ln\hat{\theta}(k,\ell) - \frac{\mu_\theta}{2}\cdot J'_\theta(k,\ell),} \tag{4.34}$$

where $\theta \in \{A,B,C\}$ denotes a model parameter and μ_θ denotes the step-size used to update it. J'_θ denotes the partial derivative of the cost function J in (4.33) w.r.t. the logarithm of the parameter estimate $\ln\hat{\theta}$, and is computed using (4.32) and (4.33) as:

$$\begin{aligned} J'_\theta(k,\ell) &= \frac{\partial Q^2_{\ln}(k,\ell)}{\partial \ln\hat{\theta}(k,\ell)} = 2\cdot Q_{\ln}(k,\ell)\cdot\frac{\partial Q_{\ln}(k,\ell)}{\partial \ln\hat{\theta}(k,\ell)} \\ &= -2\cdot\frac{Q_{\ln}(k,\ell)}{\hat{\Phi}_r(k,\ell)}\cdot\frac{\partial\hat{\Phi}_r(k,\ell)}{\partial \ln\hat{\theta}(k,\ell)}. \end{aligned} \tag{4.35}$$

Using (4.30), the partial derivative $\frac{\partial\hat{\Phi}_r(k,\ell)}{\partial \ln\hat{\theta}(k,\ell)}$ for the three model parameters is equal to:

$$\frac{\partial\hat{\Phi}_r(k,\ell)}{\partial \ln\hat{A}(k,\ell)} = \hat{A}(k,\ell)\cdot\Phi_x(k,\ell-G) + \hat{B}(k,\ell)\cdot\frac{\partial\hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln\hat{A}(k,\ell)}, \tag{4.36}$$

$$\frac{\partial\hat{\Phi}_r(k,\ell)}{\partial \ln\hat{B}(k,\ell)} = \hat{B}(k,\ell)\cdot\hat{\Phi}_{r_L}(k,\ell-1) + \hat{B}(k,\ell)\cdot\frac{\partial\hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln\hat{B}(k,\ell)}, \tag{4.37}$$

$$\frac{\partial\hat{\Phi}_r(k,\ell)}{\partial \ln\hat{C}(k,\ell)} = \hat{C}(k,\ell)\cdot\sum_{g=0}^{G-1}\Phi_x(k,\ell-g). \tag{4.38}$$

It can be observed that the right hand side of (4.36) and (4.37) contain the partial derivatives of the LRE PSD estimate in the *previous* frame $\hat{\Phi}_{r_L}(k,\ell-1)$ w.r.t. the logarithms of the parameter estimates in the *current* frame $\ln\hat{A}(k,\ell)$ and $\ln\hat{B}(k,\ell)$, respectively. These terms exist due to the recursive model for the LRE PSD in (4.24). These partial derivatives cannot be computed in a straightforward manner, as $\hat{\Phi}_{r_L}(k,\ell-1)$ does not directly depend on either $\hat{A}(k,\ell)$ or $\hat{B}(k,\ell)$. In the following subsections, we present two algorithms which have been proposed in [75] to approximate these partial derivatives.

4.5.3 *Recursive prediction error (RPE)*

The RPE adaptive algorithm approximates the partial derivatives using the parameter estimates in the previous frame:

$$\begin{aligned}\frac{\partial\hat{\Phi}_{r_L}(k,\ell-1)}{\partial\ln\hat{A}(k,\ell)} &\approx \frac{\partial\hat{\Phi}_{r_L}(k,\ell-1)}{\partial\ln\hat{A}(k,\ell-1)},\\ \frac{\partial\hat{\Phi}_{r_L}(k,\ell-1)}{\partial\ln\hat{B}(k,\ell)} &\approx \frac{\partial\hat{\Phi}_{r_L}(k,\ell-1)}{\partial\ln\hat{B}(k,\ell-1)},\end{aligned} \tag{4.39}$$

which are reasonable approximations if the step-sizes μ_A and μ_B used to update the reverberation parameters are sufficiently small. Using these approximations in (4.36) and (4.37) enable computing the partial derivatives recursively.

4.5.4 *Pseudo linear regression (PLR)*

The PLR algorithm simply assumes that the LRE PSD estimate in the previous frame $\hat{\Phi}_{r_L}(k,\ell-1)$ is independent of the parameter estimates in the current frame, i.e.

$$\frac{\partial\hat{\Phi}_{r_L}(k,\ell-1)}{\partial\ln\hat{A}(k,\ell)} = 0, \qquad \frac{\partial\hat{\Phi}_{r_L}(k,\ell-1)}{\partial\ln\hat{B}(k,\ell)} = 0. \tag{4.40}$$

It should be noted that these assumptions are stronger than the ones used for the RPE algorithm in (4.39). Using these assumptions in (4.36) and (4.37) yield *non-recursive* formulations for the partial derivatives, which are therefore approximate versions of the partial derivatives computed using the RPE algorithm.

4.6 Simulation results

In this section, we evaluate the performance of the proposed parameter estimation methods, i.e., the OE-RPE-MSLE and the OE-PLR-MSLE, and compare their performance with the state-of-the-art signal-based methods discussed in Section 4.5.1. We will refer to the proposed methods estimating all three model parameters as 3P methods, whereas we will refer to the methods in [78] estimating only the two

Parameter	Values
σ_E^2	$\{-60, -50, -40, -30, -20, -10\}$ dB
σ_L^2	$\{-40, -36, -32, -28, -24, -20\}$ dB
T_{60}	$\{200, 400, 600, 800, 1000\}$ ms

Table 4.1: Parameter values for generating the artificial IRs.

reverberation parameters as 2P methods. In Sections 4.6.1 and 4.6.2, we present the acoustic conditions and algorithmic parameters used in our simulations. In Section 4.6.3 we discuss the performance metrics used to evaluate the PSD estimation accuracy, residual echo suppression and near-end speech distortion. In Section 4.6.4, we present the simulation results for two settings: an idealistic setting using artificially generated IRs and a realistic setting using real-world IRs.

4.6.1 *Acoustic conditions*

For all simulations, the sampling frequency of the time-domain signals is equal to $f_s = 16$ kHz. The 30s long far-end signal x and the 5s long near-end speech signal s are obtained from the TIMIT database [98], where the double-talk condition occurs in the last 5s. The 30s long background noise signal v is stationary air conditioner noise measured in a quiet office.

Two different types of IRs have been considered in our simulations:

- Artificial IRs: the artificial IRs were generated according to the following time-domain model:

$$\Delta h(i) = \begin{cases} w_E(i), & 0 \leq i < N \\ w_L(i) \cdot e^{-\rho\ (i-N)}, & N \leq i < N_h, \end{cases} \tag{4.41}$$

  where $w_E \sim \mathcal{N}(0, \sigma_E^2)$ and $w_L \sim \mathcal{N}(0, \sigma_L^2)$ are zero-mean white Gaussian noise processes with variances σ_E^2 and σ_L^2, respectively, and ρ denotes the decay rate defined in (4.23). This model assumes that the first N coefficients of Δh correspond to the AEC misalignment filter (in the time-domain), where the misalignment is spread evenly over all AEC filter coefficients, whereas the later coefficients of Δh correspond to the exponentially decaying model in (4.22). The IR parameters σ_L^2 and ρ are related to the (frequency-independent) model parameters A and B as in (4.25) and (4.26), while the IR parameter σ_E^2 is related to the (frequency-independent) parameter C as $C = \sigma_E^2 \cdot F$ (see Appendix B.2). A total of 180 artificial IRs were generated using all combinations of the frequency-independent parameters σ_E^2, σ_L^2 and T_{60} given in Table 4.1, with $N = 640$ and $N_h = 16000$.

Room	No. of IRs	T_{60}	Shape
Lab	16	300-400 ms	Rectangular
Garage	16	400-500 ms	Rectangular
Office	16	500-600 ms	L-shaped
Echoic	7	850-950 ms	Rectangular

Table 4.2: Details about measured IRs.

Method	$\mu_{\mathbf{A}}$	$\mu_{\mathbf{B}}$	$\mu_{\mathbf{C}}$
OE-RPE-MSLE	10^{-2}	10^{-4}	10^{-1}
OE-PLR-MSLE	$10^{-1.5}$	$10^{-3.5}$	$10^{-0.5}$

Table 4.3: Step-sizes used for the proposed methods.

- Measured IRs: Similarly as in [78], we considered a total of 55 IRs measured in four rooms with different reverberation times, with details given in Table 4.2. The broadband T_{60} value of each IR was estimated via line-fitting on its corresponding energy decay curve [29].

4.6.2 *Algorithmic parameters*

For the subband processing, a filterbank of order $N_{\text{FFT}} = 512$ (i.e., $K = 257$) and an overlap of 75% (i.e., frameshift $F = 128$) have been used, with a Hann window as the analysis window. All required PSDs have been computed via recursive smoothing according to (4.11), using a smoothing factor $\alpha = e^{\frac{-2 \cdot F}{f_s \cdot t_c}}$ with a time-constant $t_c = 0.02$s. For the postfilter in (4.12), an over-estimation factor $\beta = 2$ and a spectral floor $\gamma = -20$ dB have been used. For the proposed methods, the step-sizes listed in Table 4.3 were found to yield good results when used in (4.34). For the state-of-the-art methods, the following parameters have been used:

- Hänsler's method [1]: smoothing factor $\delta = 0.9$ in (4.27)
- Valero's method [23]: batch size $N_T = 3750$ frames
- Modified Favrot's method (see Appendix B.1): $N = 640$, $O = 1024$ and $P = \kappa \cdot F$ with $\kappa = 12$.

4.6.3 *Performance metrics*

To evaluate the estimation accuracy of the residual echo PSD, we compute the Log Spectral Distance (LSD) [91] between the target PSD Φ_r and the residual echo PSD estimate $\hat{\Phi}_r$ in (4.30), defined as:

$$\text{LSD} = \frac{10}{K \cdot L} \cdot \sum_{k=0}^{K-1} \sum_{\ell=l_1+1}^{l_1+L} \left| \log_{10} \left(\frac{\Phi_r(k,\ell)}{\hat{\Phi}_r(k,\ell)} \right) \right|, \tag{4.42}$$

where l_1 and L denote the start and the duration of the evaluation window in frames, respectively. We choose the evaluation window to be between 20s and 25s, i.e., $l_1 = 2500$ and $L = 625$. If the LSD score is low, it means that the residual echo PSD estimate is accurate, with the perfect estimate $\hat{\Phi}_r(k,\ell) = \Phi_r(k,\ell)$ resulting in LSD $= 0$.

To evaluate the amount of residual echo suppression after applying the postfilter, we compute the segmental residual echo attenuation, defined as:

$$\text{REA}_{\text{seg}} = \frac{10}{L} \cdot \sum_{\ell=l_1+1}^{l_1+L} \log_{10} \left(\frac{\sum_{f=0}^{F-1} r^2(\ell \cdot F + f)}{\sum_{f=0}^{F-1} \tilde{r}^2(\ell \cdot F + f)} \right), \tag{4.43}$$

where the time-domain signals r and $\tilde{r}$ are obtained through inverse STFT processing of the residual echo signal R and its postfiltered version $\tilde{R}$, respectively (see Section 4.3.2). If the REA_{seg} score is high, it means that a large amount of residual echo has been suppressed, which is desirable.

Similarly, to evaluate the amount of near-end speech distortion, we compute the segmental speech-to-speech distortion ratio [99–101], defined as:

$$\text{SSDR}_{\text{seg}} = \frac{10}{L} \cdot \sum_{\ell=l_2+1}^{l_2+L} \log_{10} \left(\frac{\sum_{f=0}^{F-1} s^2(\ell \cdot F + f)}{\sum_{f=0}^{F-1} s_d^2(\ell \cdot F + f)} \right), \tag{4.44}$$

where $s_d(n) = s(n) - \tilde{s}(n)$, with $\tilde{s}$ obtained through inverse STFT processing of the postfiltered near-end speech signal $\tilde{S}$. This score is computed during periods of double-talk, which occurs between 25s and 30s, i.e., $l_2 = 3125$. If the SSDR_{seg} score is high, it means that the distortion of the near-end speech signal is low, which is desirable.

4.6.4 *Experimental results*

The first experiment is performed in an idealistic setting using artificially generated IRs, no background noise and no near-end speech. This experiment aims at evaluating the estimation accuracy of the proposed 3P methods and their simplified 2P versions in [78] for the residual echo PSD and the artificial IR parameters. The second experiment is performed in a realistic setting using real-world IRs, near-end

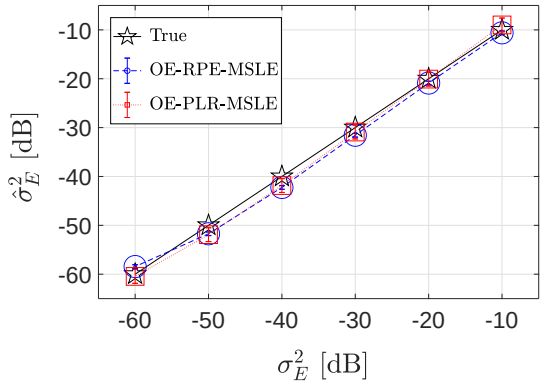

Fig. 4.5: Plot of $\hat{\sigma}_E^2$ vs. σ_E^2 for the proposed methods in the idealistic setting.

speech, background noise and a pre-converged subband AEC filter. This experiment aims at comparing the PSD estimation accuracy and the residual echo suppression performance of the proposed methods with the considered state-of-the-art methods.

4.6.4.1 *Idealistic setting*

In this experiment, we use the artificially generated IRs (see Table 4.1) to generate the acoustic echo signal and use no AEC filter, i.e., $\underline{\hat{H}}(k) = [0 \dots 0]^T$. Additionally, we assume no near-end speech ($s = 0$) and background noise ($v = 0$). This means that:

$$y(n) = d(n) = \sum_{i=0}^{N_h-1} \Delta h(i) \cdot x(n-i), \tag{4.45}$$

with Δh defined in (4.41) and $E(k,\ell) = Y(k,\ell)$. For these idealistic settings, we evaluate the accuracy of the residual echo PSD estimate $\hat{\Phi}_r$ obtained using the proposed methods and compare the estimates of the artificial IR parameters $\hat{\sigma}_E^2$, $\hat{\sigma}_L^2$ and $\hat{T}_{60}$ with the true values. These parameter estimates are obtained by averaging the converged values of $\hat{A}(k)$, $\hat{B}(k)$ and $\hat{C}(k)$ over all frequency bins and feeding them in (4.25), (4.26) and (B.9), respectively.

Fig. 4.5, 4.6 and 4.7 show the true variance of the misalignment σ_E^2 against the estimated variance $\hat{\sigma}_E^2$, the true variance of the late part of the IR σ_L^2 against the estimated variance $\hat{\sigma}_L^2$, and the true reverberation time T_{60} against the estimated reverberation time $\hat{T}_{60}$, obtained using the proposed methods, respectively. Each point in Fig. 4.5 is obtained by averaging the estimates $\hat{\sigma}_E^2$ over 30 IRs with different σ_L^2 and T_{60} values, in Fig. 4.6 by averaging the estimates $\hat{\sigma}_L^2$ over 30 IRs with different σ_E^2 and T_{60} values and in Fig. 4.7 by averaging the estimates $\hat{T}_{60}$ over 36 IRs with different σ_E^2 and σ_L^2 values, respectively. The error bars depict the standard deviations across the respective IRs. It can be observed from Fig. 4.5 that for both methods, the parameter σ_E^2 can be estimated very accurately (with

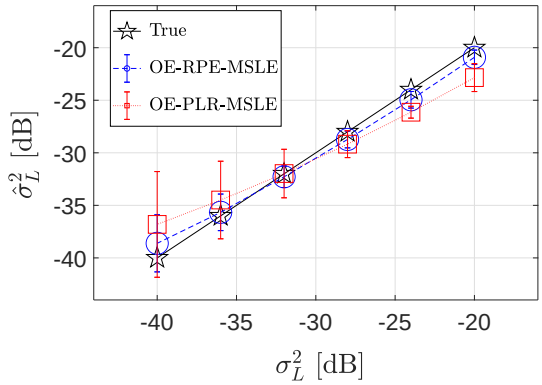

Fig. 4.6: Plot of $\hat{\sigma}_L^2$ vs. σ_L^2 for the proposed methods in the idealistic setting.

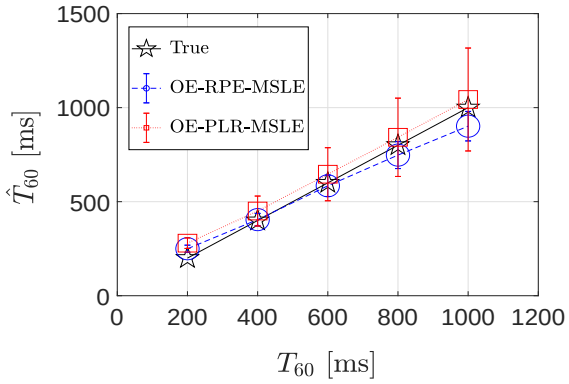

Fig. 4.7: Plot of $\hat{T}_{60}$ vs. T_{60} for the proposed methods in the idealistic setting.

very small standard deviations) over a large range of parameter values, indicating robustness to different values of σ_L^2 and T_{60}. In addition, it can be observed from Fig. 4.6 and 4.7 that the RPE algorithm typically yields more accurate estimates (and especially smaller standard deviations) of the parameters σ_L^2 and T_{60} than the PLR algorithm over a large range of parameter values. This is not surprising, since the PLR algorithm is an approximation of the RPE algorithm.

We now investigate the benefit of estimating all three model parameters using the proposed 3P methods against estimating only two model parameters using the 2P methods in [78]. To this end, we compare the influence of different amounts of misalignment, represented by σ_E^2, on the estimation accuracy of the parameters σ_L^2 and T_{60}. For $\sigma_L^2 = -32$ dB, Fig. 4.8 shows the estimated variance $\hat{\sigma}_L^2$ obtained using the 2P and 3P estimation methods for different values of σ_E^2. Each point is obtained by averaging the estimates over 6 IRs with different T_{60} values. For $T_{60} = 600$ ms, Fig. 4.9 shows the estimated reverberation time $\hat{T}_{60}$ obtained using the 2P and 3P estimation methods for different values of σ_E^2. Each point is obtained by averaging the estimates over 6 IRs with different σ_L^2 values. It should be noted that $\sigma_E^2 = -\infty$

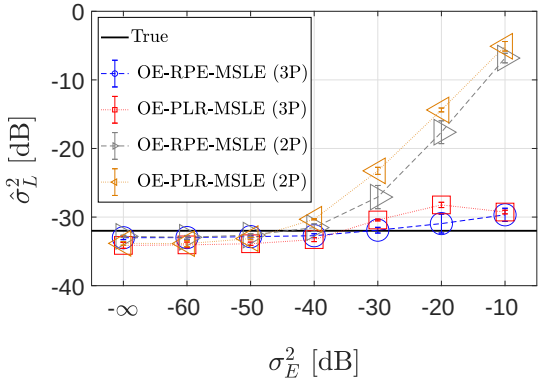

Fig. 4.8: Plot of $\hat{\sigma}_L^2$ obtained using the proposed methods (2P and 3P versions) as a function of different variances σ_E^2 in the idealistic setting ($\sigma_L^2 = -32$ dB).

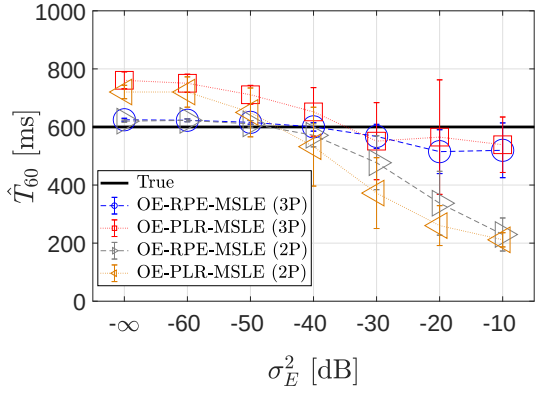

Fig. 4.9: Plot of $\hat{T}_{60}$ obtained using the proposed methods (2P and 3P versions) as a function of different variances σ_E^2 in the idealistic setting ($T_{60} = 600$ ms).

dB corresponds to no filter misalignment, i.e., no early residual echo. It can be observed that the 2P methods yield accurate estimates for the σ_L^2 and T_{60} parameters only for low values of σ_E^2, and fail to do so for large amounts of misalignment. On the other hand, the proposed 3P methods yield accurate estimates for both parameters for all considered σ_E^2 values, where the RPE algorithm again outperforms the PLR algorithm. These results clearly show the benefit of estimating all three model parameters, especially when a significant amount of filter misalignment is present.

Fig. 4.10 shows the LSD scores between the target and estimated residual echo PSDs, obtained using the 2P and 3P estimation methods for different values of σ_E^2. Each point is obtained by averaging the LSD scores over 30 IRs with different σ_L^2 and T_{60} values, while each error bar depicts the standard deviation across these IRs. It can be observed that the proposed 3P methods yield more accurate estimates for the residual echo PSD, especially for large values of σ_E^2. This again clearly shows

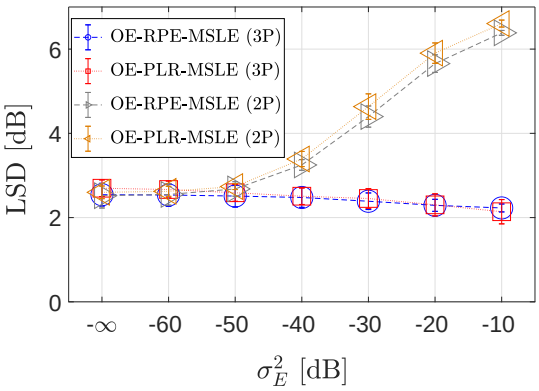

Fig. 4.10: LSD scores obtained using the proposed methods (2P and 3P versions) as a function of different variances σ_E^2 in the idealistic setting.

the benefit of estimating all three model parameters. In terms of LSD scores, it should be noted that the RPE and PLR algorithms yield similar results.

4.6.4.2 *Realistic setting*

In this experiment, we use IRs measured in different rooms (see Table 4.2) to generate the acoustic echo signal and a pre-converged[1] subband AEC filter $\underline{\hat{H}}$. The length of the AEC filter is rather short ($G = 5$, corresponding to 64 ms), covering just the direct path and early reflections in the IRs. In addition, near-end speech and background noise are present at a signal-to-noise ratio of 40 dB. In order to achieve a fair comparison between the segmental metrics for all IRs, each IR has been scaled appropriately such that a speech-to-residual echo ratio of 10 dB is obtained. The model parameters are estimated only during periods of near-end speech absence, i.e., during the first 25s, and during periods when the AEC error PSD Φ_e is at least 3 dB above the background noise PSD Φ_v. For these realistic settings, we compare the LSD, REA_{seg} and SSDR_{seg} scores obtained for the proposed methods with those obtained for the considered state-of-the-art methods.

Fig. 4.11 shows the LSD scores obtained using all considered parameter estimation methods for different rooms. Each point is obtained by averaging the LSD scores over all IRs in a room, with the error bars depicting the standard deviation across these IRs. The rooms have been placed in order of increasing T_{60} from left to right. It can be observed that both proposed methods consistently estimate the residual echo PSD more accurately than all other methods, with the next best performances delivered by Valero's and Favrot's methods. Hänsler's method, which uses just a

1 The filter was converged using white Gaussian noise as the far-end signal and the subband NLMS algorithm [15] for updating the filter coefficients.

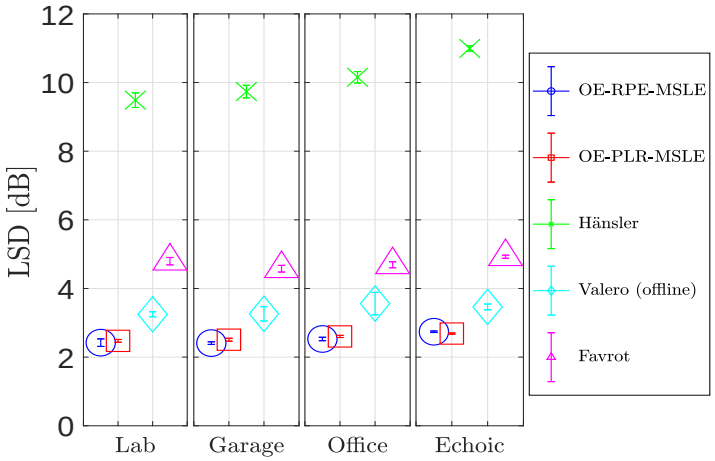

Fig. 4.11: LSD scores obtained using all considered parameter estimation methods for different rooms.

single parameter (coupling factor) to estimate the complete residual echo PSD, yields the highest LSD scores for all rooms, which is to be expected.

Fig. 4.12 shows the $\mathrm{REA}_{\mathrm{seg}}$ scores plotted against the $\mathrm{SSDR}_{\mathrm{seg}}$ scores obtained using all considered methods for different rooms. Each point is obtained by averaging the segmental metrics obtained for all IRs in a room, with the error bars on the x-axis and y-axis depicting the standard deviations across these IRs. For comparison, the scores obtained using the perfect residual echo PSD estimate $\hat{\Phi}_r(k,\ell) = \Phi_r(k,\ell)$ and an over-estimation factor $\beta = 1$ are also included, which corresponds to the best possible performance in terms of maximizing both segmental metrics. It can be observed that both proposed methods and Valero's method yield the highest $\mathrm{SSDR}_{\mathrm{seg}}$ scores (about 2-5 dB better than other methods), but the proposed methods outperform Valero's method in terms of the $\mathrm{REA}_{\mathrm{seg}}$ score. In addition, it can be observed that the proposed method with the RPE algorithm and Hänsler's method yield the highest $\mathrm{REA}_{\mathrm{seg}}$ scores (about 1-2 dB better than other methods), but the proposed method clearly outperforms Hänsler's method in terms of the $\mathrm{SSDR}_{\mathrm{seg}}$ score. In conclusion, the proposed method with the RPE algorithm provides the best performance in terms of maximizing both segmental metrics.

4.7 Conclusions

In this paper, we proposed two signal-based methods to jointly estimate the PSD of ERE and LRE components based on parametric models. We modeled the ERE PSD (due to filter misalignment) using a moving average filter on the PSD of the far-end signal and the LRE PSD (due to under-modeling of the echo path by the AEC filter) using an IIR filter on the PSD of the far-end signal. The estimated residual echo PSD was then fed into a postfilter used for residual echo suppression. The coefficients of

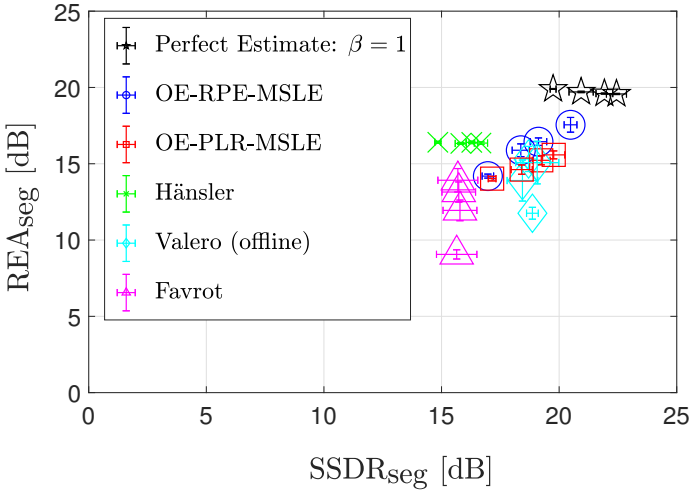

Fig. 4.12: Segmental residual echo attenuation ($\mathrm{REA_{seg}}$) vs. segmental speech-to-speech distortion ratio ($\mathrm{SSDR_{seg}}$) scores obtained using all considered parameter estimation methods for different rooms.

the moving average filter were modeled using a frequency-dependent coupling factor, while the IIR filter coefficients were modeled using frequency-dependent reverberation scaling and decay parameters. The three model parameters were estimated jointly in online mode using the signal-based output error method by minimizing a single MSLE cost function, with the parameters updated simultaneously using either the RPE or PLR algorithm. The proposed methods were first evaluated in an idealistic setting (artificially generated IRs, no near-end speech and no background noise), and yielded accurate estimates for all three model parameters and the residual echo PSD, with the RPE algorithm performing better than the PLR algorithm. Additionally, the proposed methods were compared with their simpler versions in [78], which do not model the filter misalignment and use just an IIR filter to model the complete residual echo. The proposed methods yielded accurate parameter and residual echo PSD estimates irrespective of the amount of filter misalignment present, while the simpler versions failed completely for high amounts of filter misalignment. The proposed methods were then compared with state-of-the-art parameter estimation methods in a realistic setting (IRs measured in different rooms, pre-converged AEC filter, near-end speech and background noise). The proposed method with the RPE algorithm consistently outperformed all other methods in terms of estimation accuracy of the residual echo PSD and delivered the best performance in terms of maximizing residual echo suppression while minimizing near-end speech distortion.

5

CONCLUSION AND FURTHER RESEARCH

In this chapter, we provide a summary of the main contributions of the thesis and discuss possible directions for further research.

5.1 Conclusion

Hands-free speech communication devices have become increasingly popular during the last decades and have been widely deployed for applications such as teleconferencing and voice-controlled applications. In addition to capturing the desired near-end speech from the user, these devices also pick up acoustic echo due to the acoustic coupling between the loudspeakers and microphones, which may result in a significant reduction in quality and intelligibility of the near-end speech. AEC systems are typically required in such devices, which consist of adaptive filters aiming at estimating the IRs between the loudspeakers and microphones. In reverberant environments long AEC filters are required to achieve good echo cancellation performance. This, however, results in large computational cost for the filter update as well as slow convergence. Using short AEC filters, on the other hand, results in the filters being unable to model the complete echo paths between the loudspeakers and microphones, possibly leading to a significant amount of residual echo. This residual echo is typically suppressed using a RES postfilter in the subband domain, which requires an accurate estimate of the residual echo PSD.

The main objective of this thesis was to investigate and develop tap selection schemes for implementing partial updates of multichannel AEC filters in order to achieve low-complexity AEC, as well as to improve model-based estimators for the residual echo PSD in order to achieve effective RES. In Chapter 2, we proposed novel tap selection schemes which exploit the sparsity present in the multichannel loudspeaker signals across frequency, channels and time, for partially updating subband MAEC filters. In Chapter 3, we proposed novel signal-based methods to jointly estimate the frequency-dependent reverberation scaling and decay parameters in online mode, which were used to estimate the PSD of the LRE component. In Chapter 4, we proposed a novel model for the PSD of the ERE component based on a frequency-dependent coupling factor and combined it with the LRE PSD model in Chapter 3 to yield a new model for the complete residual echo PSD. Additionally, we proposed signal-based methods to jointly estimate all three model parameters

(both reverberation parameters and the coupling factor) in online mode. These proposed methods can be seen as extensions of the methods in Chapter 3, where the ERE component was not considered.

In **Chapter 2**, we first analyzed the sparsity present in real-world multichannel signals across the dimensions of frequency, channels and time. An analysis of a 5-channel movie signal showed it to be on average about 65% sparse across subbands, 44% sparse across frames and 62% sparse across channels, respectively. However, analyzing the sparsity *jointly* across all three dimensions revealed the signal to be about 90% sparse. A similar analysis for mono speech and 5-channel music signals revealed a joint sparsity across all three dimensions of 92% and 70%, respectively. We investigated existing tap selection schemes and proposed novel schemes which exploit this high amount of input signal sparsity for partially updating subband MAEC filters, thereby saving computational cost. We first investigated the 3DM scheme, which simply applies the M-Max criterion on the tap-inputs across all three dimensions, and the SPU scheme, which only selects taps in filters with the largest magnitude tap-inputs. Simulation results for spectrally and spatially sparse synthetic signals, such as mono brown noise and stereo white noise signals, showed that the 3DM and SPU schemes completely ignore filters with the smallest magnitude tap-inputs, leading to slow filter convergence and low ERLE values. In order to overcome this problem, we proposed two tap selection schemes which do not ignore any filters for update. The FEA scheme selects the same number of filter taps in each subband and channel, thereby not exploiting input signal sparsity across frequency and channels. The DEA scheme exploits this sparsity by selecting more taps in filters with relatively larger magnitude tap-inputs, while not ignoring filters with smaller magnitude tap-inputs. Simulation results for mono speech and 5-channel music signals with only 20% of all filter taps updated in every frame showed that the 3DM and the DEA schemes perform almost as well as updating all filter taps in terms of ERLE (about 1 dB worse), with the FEA scheme performing slightly worse (about 2-4 dB) and the SPU scheme performing significantly worse (about 8-12 dB). However, even when only 20% of all filter taps are updated, the 3DM scheme still requires about 94% of the total computational effort needed for full filter update, primarily due to the large computational cost of sorting all tap-inputs in each frame. On the other hand, the SPU, FEA and DEA schemes do not need such large sorting effort and require only about 28% of the total computational effort needed for full filter update. Therefore, in conclusion, the proposed DEA tap selection scheme consistently achieves similar echo cancellation performance to full filter update at a significantly reduced computational cost for both synthetic and real-world multichannel signals.

In **Chapter 3**, we considered a single-channel system and used a subband AEC filter which is long enough to cover the direct path and early reflections of the IR, such that the LRE component contained only late reverberation. Based on a statistical reverberation model for the late reverberant part of an IR, we modeled the relationship between the LRE PSD and the PSD of the loudspeaker signal using an IIR filter with two frequency-dependent parameters, namely the reverberation scaling and decay parameters. The reverberation scaling parameter is related to

the initial power of the LRE component, while the reverberation decay parameter is related to the reverberation time T_{60}. We proposed two signal-based methods, namely the OE and EE methods, to jointly estimate both reverberation parameters by minimizing either an MSE or an MSLE cost function in online mode. For the OE method, we used gradient-descent-based algorithms such as the RPE and PLR algorithms to update the parameters. The proposed online parameter estimation methods were first evaluated in an idealistic setting using artificially generated IRs, no filter misalignment, no near-end speech and no background noise signals. Simulation results for this idealistic setting showed that the OE-RPE-MSLE method consistently outperforms all other proposed methods in terms of estimation accuracy of the reverberation parameters as well as the LRE PSD, yielding the lowest LSD scores (2.0-2.5 dB) for all considered IRs. The proposed methods were then compared with state-of-the-art offline and online parameter estimation methods in a realistic setting using IRs measured in different rooms, a fully converged AEC filter, near-end speech and background noise signals. Simulation results for this realistic setting showed that the proposed OE-RPE-MSLE method consistently outperforms state-of-the-art and the other proposed methods in terms of estimation accuracy of the T_{60} and the LRE PSD, yielding the lowest LSD scores (2.5-3.0 dB) for all considered IRs. Among all considered methods, it also yields the highest SSDR_{seg} scores (18-20 dB), with most other methods performing significantly worse (about 5-10 dB worse), while also yielding respectable REA_{seg} scores (14-16 dB), with most other methods only performing marginally better (about 2-3 dB better). Therefore, in conclusion, the proposed OE-RPE-MSLE method yields the most accurate estimates for both reverberation parameters as well as for the LRE PSD, and results in the smallest amount of near-end speech distortion while delivering a large amount of residual echo suppression.

In **Chapter 4**, we also considered the ERE component caused by the misalignment between the AEC filter and the IR. By assuming that the filter misalignment spreads evenly over all AEC filter taps, we proposed to model the ERE PSD using a moving average filter (with the same filter length as the subband AEC filter) on the PSD of the loudspeaker signal, based on a single frequency-dependent coupling factor. The proposed moving average filter model for the ERE PSD was then combined with the IIR filter model for the LRE PSD considered in Chapter 3 to yield a new model for the complete residual echo PSD. Based on the results obtained in Chapter 3, we only considered the OE method with the RPE and PLR algorithms to jointly estimate all three model parameters (both reverberation parameters and the coupling factor) by minimizing a single MSLE cost function in online mode. The proposed methods estimating the three model parameters (3P) were first evaluated in an idealistic setting using artificially generated IRs, no near-end speech and no background noise signals. Simulation results showed that the proposed methods yield accurate estimates for all three model parameters, with the RPE algorithm performing better than the PLR algorithm. In this idealistic setting, they were then compared with their simplified versions from Chapter 3, which only estimate the two reverberation parameters (2P), to illustrate the benefit of estimating all three model parameters. Simulation results showed that the proposed 3P methods provide

highly accurate estimates for all model parameters and yield low LSD scores (2.0-2.5 dB), irrespective of the amount of filter misalignment, while the 2P methods fail completely for high amounts of misalignment, yielding very high LSD scores (about 6-7 dB). The proposed 3P methods were then compared with state-of-the-art offline and online parameter estimation methods in a realistic setting using IRs measured in different rooms, a pre-converged AEC filter, near-end speech and background noise signals. Simulation results for this realistic setting showed that the proposed OE-RPE-MSLE (3P) method consistently outperforms all other methods in terms of estimation accuracy of the residual echo PSD, yielding the lowest LSD scores (2.0-2.5 dB) for all considered IRs. Among all considered methods, it also yields the highest SSDR_{seg} scores (17-21 dB), with the other methods performing about 2-5 dB worse, as well as the highest REA_{seg} scores (14-18 dB), with the other methods performing about 1-2 dB worse. Therefore, in conclusion, the proposed OE-RPE-MSLE (3P) method yields the most accurate estimates for all three model parameters as well as for the residual echo PSD, and delivers the best performance in terms of maximizing residual echo suppression while minimizing near-end speech distortion.

5.2 Further research directions

In this section, we summarize possible research directions for further improvements and potential applications of the proposed tap selection schemes for MAEC and parameter estimation methods for RES.

In Chapter 2, we designed the DEA tap selection scheme by first ranking the MAEC filters in the different subbands and channels using their respective tap-input magnitudes and then choosing specific criteria to fulfill a constraint. Alternative approaches could be used to rank the different filters, e.g., using the l_2 norm of the tap-inputs, and different criteria could be considered to design the DEA scheme. The constraint, which currently needs to be predefined, could be determined dynamically by performing sparsity analysis on the multichannel input signals and adjusted accordingly in real-time, e.g., if the input signals are found to be highly sparse, then the constraint could be lowered. Aiming to improve the filter convergence and tracking performance, the feasibility of incorporating the proposed tap selection schemes with the AP and RLS algorithms [1, 15] could be explored.

In this thesis, all simulations have been performed in rooms with static IRs and with pre-converged AEC filters. An important study could be to investigate the tracking performance of the proposed parameter estimation methods in Chapters 3 and 4 by incorporating room changes in the simulations. A future direction of research could be to develop a better understanding of the effect that the AEC filter length has on the performance of the AEC and RES systems. A related study could be to investigate the performance of the proposed parameter estimation methods when no AEC filter is present, i.e., in a pure acoustic echo suppression setup. Another study could be to investigate the performance of the proposed RES methods when the AEC filter is still converging, or when the AEC filter is updated using a tap selection scheme, e.g., using the proposed tap selections in Chapter 2.

For modeling the ERE PSD, the moving average filter in Chapter 4 could be replaced with a generic FIR filter with a large number of coefficients, with all filter coefficients modeled independently of each other. If computational complexity is not a limitation, the complete residual echo PSD could be modeled using a deep neural network (DNN) with thousands of parameters [102], with the parameters trained using supervised deep learning algorithms. Additionally, a study could be performed to compare our classical AEC filter and RES postfilter based method with end-to-end deep learning methods [103–106], where the AEC and RES systems are replaced by a single DNN, and hybrid deep learning methods, where classical approaches such as using adaptive filters for AEC are combined with deep learning approaches such as using DNNs or recurrent neural networks for RES [107–110].

Finally, the usage of the proposed tap selection schemes and parameter estimation methods could be investigated for other applications, such as acoustic feedback cancellation, active noise control, network echo cancellation etc.

A

APPENDIX FOR CHAPTER 3

A.1 Derivation of model for late residual echo PSD

We adopt the methodology used in [111] and [112] to derive the recursive expression for λ_{r_L} in (3.14), as well as expressions for the reverberation parameters A and B in terms of the RIR model parameters σ_L^2 and ρ in (3.15) and (3.16). The energy envelope of the late part of the stochastic RIR h in (3.12) is given as:

$$E\{h^2(i)\} = \sigma_L^2 \cdot e^{-2\rho(i-N)}, \;\; N \leq i < N_h, \tag{A.1}$$

where $E\{\cdot\}$ denotes spatial expectation, i.e. the ensemble average over different realizations of the stochastic process h. As the LRE signal r_L is given as:

$$r_L(n) = \sum_{i=N}^{N_h-1} h(i) \cdot x(n-i), \tag{A.2}$$

its auto-correlation at lag τ for one realization of h is defined as:

$$\begin{aligned} a_{r_L r_L}(n, n+\tau; h) &= \mathcal{E}\{r_L(n) \cdot r_L(n+\tau)\} \\ &= \sum_{i=N}^{N_h-1} \sum_{j=N}^{N_h-1} h(i) \cdot h(j) \cdot \mathcal{E}\{x(n-i) \cdot x(n-j+\tau)\} \\ &= \sum_{i=N}^{N_h-1} \sum_{j=N}^{N_h-1} h(i) \cdot h(j) \cdot a_{xx}(n-i, n-j+\tau), \end{aligned} \tag{A.3}$$

where $a_{xx}(n, n+\tau)$ denotes the auto-correlation of the far-end signal $x(n)$ at lag τ. Assuming that h and x are mutually independent, the spatial average of (A.3) over all realizations of h can be computed using (A.1) as:

$$\begin{aligned} a_{r_L r_L}(n, n+\tau) &= E\{a_{r_L r_L}(n, n+\tau; h)\} \\ &= \sum_{i=N}^{N_h-1} \sum_{j=N}^{N_h-1} E\{h(i) \cdot h(j)\} \cdot a_{xx}(n-i, n-j+\tau) \\ &= \sigma_L^2 \cdot e^{2\rho N} \cdot \sum_{i=N}^{N_h-1} e^{-2\rho i} \cdot a_{xx}(n-i, n-i+\tau), \end{aligned} \tag{A.4}$$

since $E\{h(i) \cdot h(j)\} = 0$ if $i \neq j$. Evaluating (A.4) at time instant $n - F$, with $F \ll N_h$, gives:

$$\begin{aligned} a_{r_L r_L}(n-F, n-F+\tau) &= \sigma_L^2 \cdot e^{2\rho N} \cdot \sum_{i=N}^{N_h-1} e^{-2\rho i} \cdot a_{xx}(n-F-i, n-F-i+\tau) \\ &\approx \sigma_L^2 \cdot e^{2\rho N} \cdot \sum_{i=N+F}^{N_h-1} e^{-2\rho(i-F)} \cdot a_{xx}(n-i, n-i+\tau). \end{aligned} \tag{A.5}$$

Using (A.4) and (A.5), the auto-correlation of the LRE signal $a_{r_L r_L}$ can be computed recursively as:

$$\begin{aligned} a_{r_L r_L}(n, n+\tau) = e^{-2\rho F} \cdot a_{r_L r_L}(n-F, n-F+\tau) + \\ \sigma_L^2 \cdot e^{2\rho N} \cdot \sum_{i=N}^{N+F-1} e^{-2\rho i} \cdot a_{xx}(n-i, n-i+\tau). \end{aligned} \tag{A.6}$$

If we assume the signal x to be stationary over F samples, with F the STFT frameshift, (A.6) can be rewritten as:

$$\begin{aligned} a_{r_L r_L}(n, n+\tau) = e^{-2\rho F} \cdot a_{r_L r_L}(n-F, n-F+\tau) + \\ \sigma_L^2 \cdot \left(\frac{1-e^{-2\rho F}}{1-e^{-2\rho}} \right) \cdot a_{xx}(n-N, n-N+\tau). \end{aligned} \tag{A.7}$$

Using the Wiener-Khinchin theorem, (A.7) can be expressed in terms of true PSDs as:

$$\lambda_{r_L}(k, \ell) = A \cdot \lambda_x(k, \ell - G) + B \cdot \lambda_{r_L}(k, \ell - 1), \tag{A.8}$$

where $G = \lfloor \frac{N}{F} \rfloor$ and the parameters A and B are equal to:

$$A = \sigma_L^2 \cdot \left(\frac{1-e^{-2\rho F}}{1-e^{-2\rho}} \right), \tag{A.9}$$

$$B = e^{-2\rho F}. \tag{A.10}$$

A.2 Modified version of PSD estimation method in [24]

We denote the parameters estimated using the modified version of Favrot's method [24] as $\underline{\hat{\theta}}^{\mathrm{F}}$. The parameter A^{F} corresponds to the initial power of the residual echo and is estimated as:

$$\hat{A}_N^{\mathrm{F}}(k,\ell) = \frac{\mathcal{E}\{\tilde{\Phi}_e(k,\ell)\cdot\tilde{\Phi}_{x_N}(k,\ell)\}}{\mathcal{E}\{\tilde{\Phi}_{x_N}(k,\ell)\cdot\tilde{\Phi}_{x_N}(k,\ell)\}}, \tag{A.11}$$

where $\tilde{\Phi}_{x_N}(k,\ell) = |X_N(k,\ell)|^2 - \Phi_{x_N}(k,\ell)$ and $\tilde{\Phi}_e(k,\ell) = |E(k,\ell)|^2 - \Phi_e(k,\ell)$ represent the temporal fluctuations of the PSD of the N-sample delayed far-end signal $x_N(n) = x(n-N)$ and the AEC error signal $e(n)$, respectively. The far-end signal is delayed so as to temporally align it with the LRE component in the AEC error signal. Thus, the numerator in (A.11) is the cross-correlation between the temporal fluctuations of the delayed far-end signal PSD and the AEC error PSD, while the denominator is the auto-correlation of the temporal fluctuations of the delayed far-end signal PSD. The decay rate is estimated by computing (A.11) for two different delays M and $M+P$, where M should be chosen such that $\hat{A}_M^{\mathrm{F}}$ can be associated with the late reverberant part of the RIR h and P corresponds to a delay of κ frames, i.e.:

$$\hat{B}^{\mathrm{F}}(k,\ell) = \left(\frac{\hat{A}_{M+P}^{\mathrm{F}}(k,\ell)}{\hat{A}_M^{\mathrm{F}}(k,\ell)}\right)^{1/\kappa}. \tag{A.12}$$

B

APPENDIX FOR CHAPTER 4

B.1 Original and modified versions of Favrot's method

In the original method in [24], the coupling factor C was estimated as:

$$\hat{C}_{\mathrm{F}}(k,\ell) = \frac{\mathcal{E}\left\{\tilde{\Phi}_y(k,\ell)\cdot\tilde{\Phi}_{x_M}(k,\ell)\right\}}{\mathcal{E}\left\{\tilde{\Phi}_{x_M}(k,\ell)\cdot\tilde{\Phi}_{x_M}(k,\ell)\right\}} = Z_M^y(k,\ell), \tag{B.1}$$

where $\tilde{\Phi}_{x_M}(k,\ell) = |X_M(k,\ell)|^2 - \Phi_{x_M}(k,\ell)$ and $\tilde{\Phi}_y(k,\ell) = |Y(k,\ell)|^2 - \Phi_y(k,\ell)$ represent the temporal fluctuations of the PSDs of the M-sample delayed far-end signal $x_M(n) = x(n-M)$ and the microphone signal y, respectively. Here, Z_M^y is used to denote the ratio in (B.1) computed using the signals x_M and y. The delay $M(\ll N)$ was chosen so as to align the far-end signal x with the microphone signal y, i.e., it corresponds to the initial peak in the IR, which depends on the distance between the loudspeaker and the microphone. The decay rate B was estimated using (B.1) for two different signal delays O and $O+P$:

$$\hat{B}_{\mathrm{F}}(k,\ell) = \left(\frac{Z_{O+P}^y(k,\ell)}{Z_O^y(k,\ell)}\right)^{1/\kappa}, \tag{B.2}$$

where O corresponds to the late echo tail ($O \geq N$) and $P = \kappa\cdot F$ corresponds to a delay of κ frames.

The modification to the original method that we consider in this paper corresponds to using the temporal fluctuations of the AEC error PSD $\tilde{\Phi}_e(k,\ell)$ instead of $\tilde{\Phi}_y(k,\ell)$ to estimate the parameters B and C:

$$\begin{aligned}\hat{C}_{\mathrm{F}}(k,\ell) &= Z_M^e(k,\ell)\\ \hat{B}_{\mathrm{F}}(k,\ell) &= \left(\frac{Z_{O+P}^e(k,\ell)}{Z_O^e(k,\ell)}\right)^{1/\kappa}.\end{aligned} \tag{B.3}$$

Additionally, to estimate the parameter A, we use the N-sample delayed far-end signal x_N:

$$\hat{A}_{\mathrm{F}}(k,\ell) = Z_N^e(k,\ell). \tag{B.4}$$

B.2 Coupling factor

Since the ERE signal $r_E(n)$ is equal to:

$$r_E(n) = \sum_{i=0}^{N-1} \Delta h(i) \cdot x(n-i), \tag{B.5}$$

the auto-correlation for lag τ is given as:

$$\begin{aligned} a_{r_E r_E}(n, n+\tau) &= \mathcal{E}\{r_E(n) \cdot r_E(n+\tau)\} \\ &= \mathcal{E}\left\{ \sum_{i=0}^{N-1} \Delta h(i) \cdot x(n-i) \cdot \sum_{j=0}^{N-1} \Delta h(j) \cdot x(n-j+\tau) \right\} \\ &= \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \mathcal{E}\{\Delta h(i) \cdot \Delta h(j)\} \cdot a_{xx}(n-i, n-j+\tau), \end{aligned} \tag{B.6}$$

where a_{xx} denotes the auto-correlation of x. Using (4.41) and assuming that the far-end signal x is stationary over a short period of F samples, with $F \ll N$ $(= G \cdot F)$, we can rewrite (B.6) as:

$$\begin{aligned} a_{r_E r_E}(n, n+\tau) &= \sigma_E^2 \cdot \sum_{i=0}^{N-1} a_{xx}(n-i, n-i+\tau), \\ &= \sigma_E^2 \cdot \sum_{g=0}^{G-1} \sum_{f=0}^{F-1} a_{xx}(n-g \cdot F-f, n-g \cdot F-f+\tau) \\ &\approx \sigma_E^2 \cdot F \cdot \sum_{g=0}^{G-1} a_{xx}(n-g \cdot F, n-g \cdot F+\tau). \end{aligned} \tag{B.7}$$

Applying the Wiener-Khinchin theorem to (B.7) yields:

$$\lambda_{r_E}(k, \ell) = \sigma_E^2 \cdot F \cdot \sum_{g=0}^{G-1} \lambda_x(k, \ell-g), \tag{B.8}$$

such that comparing (B.8) with (4.20) yields:

$$C = \sigma_E^2 \cdot F. \tag{B.9}$$

BIBLIOGRAPHY

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. New York, NY, USA: Wiley, 2004.

[2] J. Benesty, T. Gansler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. New York, NY, USA: Springer, 2001.

[3] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*. Wiley, 2008.

[4] J. Benesty, M. M. Sondhi, and Y. Huang, *Handbook of Speech Processing*. Berlin, Germany: Springer, 2008.

[5] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech Processing in Modern Communication*. Berlin, Germany: Springer, 2010.

[6] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition", *Speech Communication*, vol. 25, no. 1–3, pp. 75–95, 1998.

[7] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners", *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, 2006.

[8] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition", *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[9] M. M. Sondhi and D. A. Berkeley, "Silencing echoes on the telephone network", in *Proc. IEEE*, vol. 68, no. 8, pp. 948–963, 1980.

[10] C. Breining et.al., "Acoustic echo control - An application of very-high-order adaptive filters", *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42–69, 1999.

[11] G. Schmidt, "Applications of acoustic echo control: an overview", in *Proc. of the European Signal Processing Conference*, Vienna, Austria, pp. 9–16, 2004.

[12] C. Faller and J. Chen, "Suppressing acoustic echo in a sampled auditory envelope space", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1048–1062, 2005.

[13] G. Enzner, "A model-based optimum filtering approach to acoustic echo control: Theory and practice", Ph.D. dissertation, RWTH Aachen University, Aachen, Germany, 2006.

[14] E. A. P. Habets, I. Cohen, S. Gannot, and P. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1433–1451, 2008.

[15] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ, USA: Prentice Hall, 1996.

[16] S. M. Kuo and J. Chen, "Multiple microphone acoustic echo cancellation system with the partial adaptive process", *Digital Signal Processing*, vol. 3, no. 1, pp. 54–63, 1993.

[17] S. C. Douglas, "Adaptive filters employing partial updates", *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 44, no. 3, pp. 209–216, 1997.

[18] T. Aboulnasr and K. Mayyas, "Complexity reduction of the NLMS algorithm via selective coefficient update", *IEEE Transactions on Signal Processing*, vol. 47, no. 5, pp. 1421–1424, 1999.

[19] K. Doğançay and O. Tanrikulu, "Adaptive filtering algorithms with selective partial updates", *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 48, no. 8, pp. 762–769, 2001.

[20] K. Doğançay and P. A. Naylor, "Recent advances in partial update and sparse adaptive filters", in *Proc. of the European Signal Processing Conference*, Antalya, Turkey, pp. 1–4, 2005.

[21] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals", *Signal Processing*, vol. 64, no. 1, pp. 33–47, 1998.

[22] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction", *IEEE Transactions on Speech Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[23] M. Valero, E. Mabande, and E. A. P. Habets, "Signal-based late residual echo spectral variance estimation", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, pp. 5914–5918, 2014.

[24] A. Favrot, C. Faller, and F. Küch, "Modeling late reverberation in acoustic echo suppression", in *Proc. of the International Workshop on Acoustic Signal Enhancement*, Aachen, Germany, pp. 1–4, 2012.

[25] H. Kuttruff, *Room Acoustics*. London, U.K.: Spon Press, 2000.

[26] W. C. Sabine, *Collected Papers on Acoustics*. Harward University Press, 1922.

[27] P. A. Naylor and N. Gaubitch, *Speech Dereverberation*. London, U.K.: Springer, 2010.

[28] J. Polack, "La transmission de l'énergie sonore dans les salles", Dissertation, Université du Maine, France, 1988.

[29] M. Schroeder, "New method of measuring reverberation time", *Journal of the Acoustical Society of America*, vol. 37, pp. 409–412, 1965.

[30] S. C. Douglas, "Analysis and implementation of the max-NLMS adaptive filter", in *Proc. of the Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, vol. 1, pp. 659–663, 1995.

[31] T. Aboulnasr and K. Mayyas, "Selective coefficient update of gradient-based adaptive algorithms", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, vol. 3, pp. 1929–1932, 1997.

[32] T. Schertler, "Selective block update of NLMS type algorithms", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, USA, vol. 3, pp. 1717–1720, 1998.

[33] O. Tanrikulu and K. Doğançay, "Selective-partial-update proportionate normalized least-mean-squares algorithm for network echo cancellation", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, vol. 2, pp. 1889–1892, 2002.

[34] S. L. Gay, "An efficient, fast converging adaptive filter for network echo cancellation", in *Proc. of the Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, vol. 1, pp. 394–398, 1998.

[35] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers", *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 508–518, 2000.

[36] J. Benesty and S. L. Gay, "An improved PNLMS algorithm", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, vol. 2, pp. 1881–1884, 2002.

[37] J. Cui, P. A. Naylor, and D. T. Brown, "An improved IPNLMS algorithm for echo cancellation in packet-switched networks", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, vol. 4, pp. 141–144, 2004.

[38] H. Deng and M. Doroslovački, "New sparse adaptive algorithms using partial update", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, vol. 2, pp. 845–848, 2004.

[39] H. Deng and M. Doroslovacki, "Improving convergence of the PNLMS algorithm for sparse impulse response identification", *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 181–184, 2005.

[40] A. W. H. Khong and P. A. Naylor, "Efficient use of sparse adaptive filters", in *Proc. of the Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, pp. 1375–1379, 2006.

[41] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation - an overview of the fundamental problem", *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, 1995.

[42] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation", *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, 1998.

[43] A. W. H. Khong, J. Benesty, and P. A. Naylor, "Stereophonic acoustic echo cancellation: analysis of the misalignment in the frequency domain", *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 33–36, 2006.

[44] A. Gilloire and V. Turbin, "Using auditory properties to improve the behavior of stereophonic acoustic echo cancellers", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, USA, vol. 6, pp. 3681–3684, 1998.

[45] S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Detroit, MI, USA, vol. 5, pp. 3059–3062, 1995.

[46] M. Ali, "Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, USA, vol. 6, pp. 3689–3692, 1998.

[47] A. W. H. Khong and P. A. Naylor, "Reducing inter-channel coherence in stereophonic acoustic echo cancellation using partial update adaptive filters", in *Proc. of the European Signal Processing Conference*, Vienna, Austria, pp. 405–408, 2004.

[48] A. W. H. Khong and P. A. Naylor, "A family of selective-tap algorithms for stereo acoustic echo cancellation", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, vol. 3, pp. 133–136, 2005.

[49] A. W. H. Khong and P. A. Naylor, "Stereophonic acoustic echo cancellation employing selective-tap adaptive algorithms", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 785–796, 2006.

[50] M. Bekrani, A. W. H. Khong, and M. Lotfizad, "A clipping-based selective-tap adaptive filtering approach to stereophonic acoustic echo cancellation", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1826–1836, 2011.

[51] J. J. Shynk, "Frequency-domain and multirate adaptive filtering", *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, 1992.

[52] K. A. Lee, W. S. Gan, and S. M. Kuo, *Subband adaptive filtering: theory and implementation*. Wiley, 2009.

[53] E. Ferrara, "Fast implementations of LMS adaptive filters", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 474–475, 1980.

[54] D. Mansour and A. H. Gray, "Unconstrained frequency-domain adaptive filter", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 5, pp. 726–734, 1982.

[55] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series", *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.

[56] J. S. Soo and K. Pang, "Multidelay block frequency domain adaptive filter", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[57] X. Lin, A. W. H. Khong, M. Doroslovački, and P. A. Naylor, "Frequency-domain adaptive algorithm for network echo cancellation in VoIP", *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, pp. 1–9, 2008.

[58] A. W. H. Khong, P. A. Naylor, and J. Benesty, "A low delay and fast converging improved proportionate algorithm for sparse system identification", *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. 1–8, 2007.

[59] P. Loganathan, X. Lin, A. W. H. Khong, and P. A. Naylor, "Frequency-domain adaptive multidelay algorithm with sparseness control for acoustic echo cancellation", in *Proc. of the European Signal Processing Conference*, Glasgow, Scotland, pp. 2002–2006, 2009.

[60] R. Crochiere, L. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ, USA: Prentice Hall, 1983.

[61] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.

[62] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three postfiltering algorithms for residual acoustic echo reduction", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, vol. 1, pp. 307–310, 1997.

[63] C. Lee, J. Shin, and N. Kim, "DNN-based residual echo suppression", in *Proc. of the Conference of the International Speech Communication Association*, Dresden, Germany, pp. 1775–1779, 2015.

[64] I. Schalk-Schupp, F. Faubel, M. Buck, and A. Wendemuth, "Approximation of a nonlinear distortion function for combined linear and nonlinear residual echo suppression", in *Proc. IEEE International Workshop on Acoustic Signal Enhancement*, Xi'an, China, pp. 1–5, 2016.

[65] I. Schalk-Schupp, F. Faubel, M. Buck, and A. Wendemuth, "Combined linear and nonlinear residual echo suppression using a deficient distortion model - a proof of concept", in *Proc. of the ITG Symposium on Speech Communication*, Paderborn, Germany, pp. 1–5, 2016.

[66] J. Franzen, and T. Fingscheidt, "An efficient residual echo suppression for multi-channel acoustic echo cancellation based on the frequency-domain adaptive Kalman filter", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, Canada, pp. 226–230, 2018.

[67] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-input neural network-based residual echo suppression", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, Canada, pp. 231–235, 2018.

[68] H. Huang, C. Hofmann, W. Kellermann, J. Chen, and J. Benesty, "A multiframe parametric Wiener filter for acoustic echo suppression", in *Proc. of the IEEE International Workshop on Acoustic Signal Enhancement*, Xi'an, China, pp. 1–5, 2016.

[69] A. Favrot, C. Faller, M. Kallinger, F. Küch, and M. Schmidt, "Acoustic echo control based on temporal fluctuations of short-time spectra", in *Proc. of the International Workshop on Acoustic Echo and Noise Control*, Seattle, WA, USA, pp. 1–4, 2008.

[70] S. Yamamoto and S. Kitayama, "An adaptive echo canceller with variable step gain method", *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 65, pp. 1–8, 1982.

[71] F. Lindstrom, C. Schüldt, and I. Claesson, "An improvement of the two-path algorithm transfer logic for acoustic echo cancellation", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1320–1326, 2007.

[72] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model", *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–774, 2009.

[73] N. K. Desiraju, S. Doclo, and T. Wolff, "Efficient multichannel acoustic echo cancellation using constrained tap selection schemes in the subband domain", *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, pp. 63–78,

2017.

[74] N. K. Desiraju, S. Doclo, T. Gerkmann, and T. Wolff, "Efficient multi-channel acoustic echo cancellation using constrained sparse filter updates in the subband domain", in *Proc. of the ITG Symposium on Speech Communication*, Erlangen, Germany, pp. 1–4, 2014.

[75] J. J. Shynk, "Adaptive IIR filtering", *IEEE ASSP Magazine*, vol. 6, no. 2, pp. 4–21, 1989.

[76] Y. Tomita, A. Damen, and P. Van Den Hof, "Equation error versus output error methods", *Ergonomics*, vol. 35, nos. 5/6, pp. 551–564, 1992.

[77] N. K. Desiraju, S. Doclo, M. Buck, T. Gerkmann, and T. Wolff, "On determining optimal reverberation parameters for late residual echo suppression", in *Proc. AES Conference on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, pp. 1–8, 2016.

[78] N. K. Desiraju, S. Doclo, M. Buck, and T. Wolff, "Online estimation of reverberation parameters for late residual echo suppression", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 77–91, 2019.

[79] N. K. Desiraju, S. Doclo, M. Buck, and T. Wolff, "Joint online estimation of early and late residual echo PSD for residual echo suppression", submitted to *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

[80] H. Buchner, J. Benesty, and W. Kellermann, "Multichannel frequency-domain adaptive filtering with application to acoustic echo cancellation", in *Adaptive signal processing: Application to real-world problems*, J. Benesty and Y. Huang, Eds., Berlin, Germany: Springer, 2003, pp. 95–128.

[81] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication", *Signal Processing*, vol. 85, no. 3, pp. 549–570, 2005.

[82] Y. Huang, J. Benesty, and J. Chen, "Identification of acoustic MIMO systems: challenges and opportunities", *Signal Processing*, vol. 86, no. 6, pp. 1278–1295, 2006.

[83] P. A. Naylor and W. Sherliker, "A short-sort M-Max NLMS partial-update adaptive filter with application to echo cancellation", in *Proc. of the IEEE International Conference on Acoustics, Speech, Signal Processing*, Hong Kong, pp. 373–376, 2003.

[84] J. M. P. Borrallo and M. G. Otero, "On the implementation of a partitioned block frequency domain adaptive filter (PBFDAF) for long acoustic echo cancellation", *Signal Processing*, vol. 27, no. 3, pp. 301–315, 1992.

[85] D. E. Knuth, *The Art of Computer Programming*, vol. 3. Reading, MA, USA: Addison-Wesley, 1973.

[86] I. Pitas, "Fast algorithms for running ordering and max/min calculation", *IEEE Transactions on Circuits and Systems*, vol. 36, no. 6, pp. 795–804, 1989.

[87] N. P. Hurley and S. T. Rickard, "Comparing measures of sparsity", *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[88] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints", *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[89] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[90] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[91] T. Gerkmann and R.C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[92] H. Schepker and S. Doclo, "Least-squares estimation of the common pole-zero filter of acoustic feedback paths in hearing aids", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 8, pp. 1334–1347, 2016.

[93] H. Schepker and S. Doclo, "A semidefinite programming approach to min-max estimation of the common part of acoustic feedback paths in hearing aids", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 2, pp. 366–377, 2016.

[94] T. Soderstrom and P. Stoica, "Some properties of the output error method", *Automatica*, vol. 18, no. 1, pp. 93–99, 1982.

[95] M. Nayeri, "A weaker sufficient condition for the unimodality of error surfaces associated with exactly matching adaptive IIR filters", in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, vol. 1, pp. 35–38, 1988.

[96] S. Stearns, "Error surfaces of recursive adaptive filters", *IEEE Transactions on Circuits and Systems*, vol. 28, no. 6, pp. 603–606, 1981.

[97] T. Soderstrom, "On the uniqueness of maximum likelihood identification", *Automatica*, vol. 11, no. 2, pp. 193–197, 1975.

[98] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM*, National Institute of Standards and Technology, 1990.

[99] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model", *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[100] T. Fingscheidt and S. Suhadi, "Quality assessment of speech enhancement systems by separation of enhanced speech, noise, and echo", in *Proc. Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, pp. 818–821, 2007.

[101] T. Matheja, M. Buck, and T. Fingscheidt, "A dynamic multi-channel speech enhancement system for distributed microphones in a car environment", *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 191, 2013.

[102] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression", in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, pp. 1-4, 2013.

[103] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions", in *Proc. of the Conference of the International Speech Communication Association*, Graz, Austria, pp. 4255–4259, 2019.

[104] A. Fazel, M. El-Khamy, and J. Lee, "CAD-AEC: Contextaware deep acoustic echo cancellation", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, pp. 6919–6923, 2020.

[105] M. M. Halimeh and W. Kellermann, "Efficient multichannel nonlinear acoustic echo cancellation based on a cooperative strategy", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, pp. 461–465, 2020.

[106] N. L. Westhausen and B. Meyer, "Acoustic echo cancellation with the dual-signal transformation LSTM network", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, pp. 7138–7142, 2021.

[107] L. Ma, H. Huang, P. Zhao, and T. Su, "Acoustic echo cancellation by combining adaptive digital filter and recurrent neural network", *arXiv preprint arXiv:2005.09237*, 2020.

[108] M. M. Halimeh, T. Haubner, A. Briegleb, A. Schmidt, and W. Kellermann, "Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, pp. 121–125, 2021.

[109] Z. Wang, Y. Na, Z. Liu, B. Tian, and Q. Fu, "Weighted recursive least square filter and neural network based residual echo suppression for the AEC-Challenge", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, pp. 141–145, 2021.

[110] R. Peng, L. Cheng, C. Zheng, and X. Li, "ICASSP 2021 Acoustic Echo Cancellation Challenge: Integrated adaptive echo cancellation with time alignment and deep learning-based residual echo plus noise suppression", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, pp. 146–150, 2021.

[111] K. Lebart and J. Boucher, "A new method based on spectral subtraction for speech dereverberation", *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[112] E. A. P. Habets, "Speech dereverberation using statistical reverberation models", in *Speech Dereverberation*, P. A. Naylor and N. Gaubitch, Eds., London, U.K: Springer, 2010, pp. 57–93.

LIST OF PUBLICATIONS

The following publications are related to the work in this thesis.

Peer-reviewed Journal Papers

[J3] N. K. Desiraju, S. Doclo, M. Buck, and T. Wolff, "Joint online estimation of early and late residual echo PSD for residual echo suppression", submitted to *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

[J2] N. K. Desiraju, S. Doclo, M. Buck, and T. Wolff, "Online estimation of reverberation parameters for late residual echo suppression", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 77–91, 2020.

[J1] N. K. Desiraju, S. Doclo, and T. Wolff, "Efficient multichannel acoustic echo cancellation using constrained tap selection schemes in the subband domain", *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, pp. 63–78, 2017.

Peer-reviewed Conference Papers

[C2] N. K. Desiraju, S. Doclo, M. Buck, T. Gerkmann, and T. Wolff, "On determining optimal reverberation parameters for late residual echo suppression", in *Proc. AES Conference on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, pp. 1–8, 2016.

[C1] N. K. Desiraju, S. Doclo, T. Gerkmann, and T. Wolff, "Efficient multi-channel acoustic echo cancellation using constrained sparse filter updates in the subband domain", in *Proc. of the ITG Symposium on Speech Communication*, Erlangen, Germany, pp. 1–4, 2014.

Patents

[P3] I. Schalk-Schupp, F. Faubel, M. Buck, N. K. Desiraju, and T. Wolff, "System and method for combined nonlinear and late echo suppression", US20190102108, Apr. 4, 2019.

[P2] T. Wolff and N. K. Desiraju, "Spectral estimation of room acoustic parameter", WO2017160294, Sep. 21, 2017.

[P1] M. Buck, T. Wolff, and N. K. Desiraju, “Residual interference suppression”, WO2016039765, Mar. 17, 2016.