Multi Channel Audio Processing: Enhancement, Compression and Evaluation of Quality

Mehrkanalige Audiosignalverarbeitung: Verbesserung, Codierung und Qualitätsbewertung

Von der Fakultät für Elektrotechnik und Informationstechnik der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Ingenieurwissenschaften genehmigte Dissertation

Diplom-Ingenieur

Magnus Schäfer

aus Arnsberg, Deutschland

Berichter: Universitätsprofessor Dr.-Ing. Peter Vary
Universitätsprofessor Dr.-Ing. Jens-Rainer Ohm

Tag der mündlichen Prüfung: 28.04.2014

Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.

AACHENER BEITRÄGE ZU DIGITALEN NACHRICHTENSYSTEMEN

Herausgeber:

Prof. Dr.-Ing. Peter Vary
Institut für Nachrichtengeräte und Datenverarbeitung
Rheinisch-Westfälische Technische Hochschule Aachen
Muffeter Weg 3a
52074 Aachen
Tel.: 0241-80 26 956
Fax.: 0241-80 22 186

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar

1. Auflage Aachen:
Wissenschaftsverlag Mainz in Aachen
(Aachener Beiträge zu digitalen Nachrichtensystemen, Band 39)
ISSN 1437-6768
ISBN 978-3-95886-007-0

Wissenschaftsverlag Mainz
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 87 34 34
Fax: 02 41 / 87 55 77
www.Verlag-Mainz.de

Herstellung: Druckerei Mainz GmbH,
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 87 34 34
www.druckereimainz.de

Gedruckt auf chlorfrei gebleichtem Papier

D 82 (Diss. RWTH Aachen University, 2014)

Acknowledgments

The research for this thesis has been done during my time as a research assistant at the Institute of Communication Systems and Data Processing (ind) at RWTH Aachen University.

I would like to express my sincere gratitude to my supervisor Prof. Dr.-Ing. Peter Vary for the valuable discussions, the continuous support and his guidance over the years. I would also like to thank Prof. Dr.-Ing. Jens-Rainer Ohm for his interest in my thesis and for being the second reader.

In addition, I would like to thank everyone that contributed to this thesis by discussions, collaborate work, proof-reading, and so much more in an always productive and pleasant working environment at the institute. My thanks go out to Andreas, Andreas, Annika, Aulis, Bastian, Benedikt, Bernd, Birgit, Carsten, Christiane, Christoph, Daniel, Florian, Frank, Hauke, Helge, Heiner, Kim, Laurent, Marco, Markus, Matthias, Matthias, Max, Moritz, Roswitha, Simone, Stefan, Sylvia, Thomas, Thomas, Tim, Tobias, and all the students that I worked with during the time at the institute.

Finally, I would like to express my deep gratitude to my family and my friends for their support and encouragement. Dear Hannah, thank you for your love, your support and your understanding – Moritz, you enjoyed my doctoral cap almost as much as I did.

Aachen, October 2014 Magnus Schäfer

Abstract

The demand for high definition audio and video services is rapidly increasing. Two representative examples for this are audio-visual conferencing or video storage and delivery. In this context, efficient techniques are required for enhancement and compression of multi channel audio signals with compatibility to, e.g., mono or stereo systems. In this thesis, novel signal processing algorithms for both enhancement and compression of multi channel signals are developed and theoretical performance bounds are derived. Additionally, a novel instrumental quality measure for the evaluation of multi channel signal processing algorithms is proposed.

Enhancement schemes for both the recording and the reproduction side are introduced. This includes the optimization of a near field filter-and-sum beamformer to achieve a target directivity characteristic at the recording side. For the reproduction side, an efficient postfilter is presented which increases the speech intelligibility by taking the positive influence of early room reflections into account.

The main part of this thesis covers multi channel predictive *compression* of audio signals. A predictive multi channel coding system is presented and analyzed. Performance bounds are derived and two methods for an adaptive bit rate distribution between inter channel and intra channel prediction are devised. Novel multi channel noise shaping concepts are introduced. The performance of the compression system is quantified by instrumental measures.

A novel instrumental measure is introduced for the *evaluation* of multi channel signal enhancement and compression. It combines the proven single channel quality measure PEAQ with a binaural auditory model and a mathematical model of cognitive behavior, providing a reliable evaluation of quality perception and spatial fidelity. The inclusion of spatial information into the instrumental quality measurement leads to a consistently high correlation between the instrumental measure and a listening test.

Zusammenfassung

Der Bedarf an hochqualitativen Audio- und Video-Diensten steigt rapide an. Beispiele hierfür sind audiovisuelle Konferenzsysteme oder Videospeicherung und -übertragung. Dafür sind effiziente Techniken für die Verbesserung und Kompression mehrkanaliger Audiosignale notwendig, wobei Kompatibilität zu Mono- oder Stereosystemen wünschenswert ist. In dieser Arbeit werden neuartige Signalverarbeitungsalgorithmen für Verbesserung und Kompression mehrkanaliger Audiosignale entwickelt und theoretische Leistungsgrenzen abgeleitet. Zusätzlich wird auch ein neues instrumentelles Qualitätsmaß für die Evaluation mehrkanaliger Signalverarbeitungsalgorithmen vorgeschlagen.

Signalverbesserungsverfahren werden sowohl für die Aufnahme- als auch für die Wiedergabeseite eingeführt. Auf der Aufnahmeseite beinhaltet dies ein Beamformingsystem, das es auf der Basis einer numerischen Optimierung erlaubt, eine Zielrichtcharakteristik zu approximieren. Auf der Wiedergabeseite wird ein effizientes Postfilter vorgestellt, das zu einer erhöhten Sprachverständlichkeit führt.

Der Hauptteil der Arbeit behandelt Systeme für die *Kompression* von Audiosignalen durch mehrkanalige lineare Prädiktion. Leistungsgrenzen der Systeme werden abgeleitet und zwei Methoden für eine adaptive Bitratenverteilung zwischen Intrakanal- und Interkanalprädiktion vorgestellt. Konzepte für den Einsatz von Rauschfärbung werden eingeführt und die Leistungsfähigkeit des kompletten Kompressionssystems wird durch instrumentelle Maße quantifiziert.

Ein neues instrumentelles Qualitätsmaß für die *Evaluation* mehrkanaliger Signalverarbeitungsalgorithmen wird eingeführt. Es kombiniert das für einkanalige Systeme bewährte Qualitätsmaß PEAQ mit einem binauralen Gehörmodell und einem mathematischen Modell der kognitiven Verarbeitung. Die Integration der räumlichen Information in die Qualitätsbewertung führt zu einer konsistent hohen Korrelation zwischen dem Maß und einem Hörversuch.

Contents

Acronyms

AAC-ELD Advanced Audio Coding – Enhanced Low Delay

AIR Aachen Impulse Response

AMR-WB Adaptive Multi-Rate Wideband

AODG Advanced Objective Difference Grade

BRIR Binaural Room Impulse Response

CEF Coherence Estimate Function

CoVR Connected Visual Reality

DCR Degradation Category Rating

DRR Direct-to-Reverberant Energy Ratio

DWT Discrete Wavelet Transform

FIR Finite Impulse Response

GMM Gaussian Mixture Model

HARQ Hybrid Automatic Repeat reQuest

HRTF Head-Related Transfer Function

IIR Infinite Impulse Response

ILD Interaural Level Difference

IP Internet Protocol

IR Impulse Response

ITD Interaural Time Difference

LisTEn LIStening Test ENvironment

LP Linear Prediction

LSF Line Spectral Frequency

LSP Line Spectral Pair

LTE Long Term Evolution

LTP Long Term Prediction

MARDY Multichannel Acoustic Reverberation Database at York

MOV Model Output Variable

MLS Maximum Length Sequence

MMSE Minimum Mean Square Error

MPEG Moving Picture Experts Group

MP3 MPEG-1 layer 3

MSC Magnitude Squared Coherence

NAP Neural Activity Pattern

NELE Near End Listening Enhancement

NN Neural Network

ODG Objective Difference Grade

PESQ Perceptual Evaluation of Speech Quality

PEAQ Perceptual Evaluation of Audio Quality

POLQA Perceptual Objective Listening Quality Assessment

PSEQ Perfect Sequence

PSD Power Spectral Density

RIR Room Impulse Response

SII Speech Intelligibility Index

SNR Signal-to-Noise Ratio

STI Speech Transmission Index

SUT System under Test

TF Transfer Function

USAC MPEG Unified Speech and Audio Coding

VoIP Voice over IP

1

Introduction

While Epictetus famously stated "Nature hath given men one tongue but two ears, that we may hear from others twice as much as we speak.", one could also argue that the two ears are only there to facilitate the sophisticated dual channel signal processing that our hearing system is capable of. Throughout this thesis, the focus is on methods for multi channel signal processing that allow to exploit or preserve the spatial properties of audio signals by means of specific signal processing algorithms.

It is well known from research on human communication that being able to localize sounds and to focus our hearing on a spatial region allows us humans to communicate even in very adverse environments. The term *cocktail party effect* was coined in [Che53] to illustrate these capabilities by the picture of people talking at a crowded cocktail party.

While there are approaches to emulate these feats of the human hearing system in the area of source separation, no system is capable of replicating every aspect yet. However, the question has to be raised if this is necessary at all since the signals that are produced by any signal processing system will be received by a human listener in the end. Hence the target for any system shall be to allow the human hearing system to work as unaffectedly as possible.

How this target can be achieved depends strongly on the application scenario. There are some common elements that are necessary, though.

- Some spatial properties have to be linked to the audio signal(s) – these properties can either be from a recording with multiple microphones or from a purely artificial rendering
- A multi channel transmission system to transmit both the audio information and the spatial properties to the receiver

- A loudspeaker setup that allows to correctly reproduce the recorded (or artificially constructed) spatial properties

Within this thesis, a two stage system is presented that contains novel methods for exploiting and transmitting spatial properties of audio signals. The outer stage is used on the transmitting side to either perform beamforming or to generate a specific mixture of the microphone signals. On the receiving side, the mixing process from the transmitting side is inverted. Additionally, a method to improve speech intelligibility by a postfiltering procedure is incorporated in this stage. The inner stage is a multi channel predictive coding scheme that is shown to allow for an efficient transmission. Special focus in this area is put on the aspect of noise shaping which is an important part of predictive coding systems as it allows to match the effective quantization error at the output to whatever target is set for the specific system.

The practical relevance of multi channel signal processing can be expected to grow in the coming years. Already, mobile devices equipped with multiple microphones have entered the market in recent years, multiple loudspeakers are also readily available in the form of stereo head sets. From the hardware side, it seems as if all the prerequisites for setting up a multi channel audio link are there. However, the microphone positions are mostly tailored to noise reduction tasks (e.g., for the system devised in [JHN+12]) and not to transmitting multiple audio signals. Once the end user advantage of having spatial information as well has been demonstrated, it would be only a question of time before the first devices appear that are optimized for this application scenario.

Even if this scenario does not materialize, there is also the area of video conferencing systems which can achieve a significantly higher perceived quality if the spatial properties of the audio signals are at least conserved if not enhanced by the transmission system. Some of the work that is presented in this thesis was carried out within the *Connected Visual Reality* (CoVR) [SHS+13] project that aimed at improving many aspects of video conferencing, namely video coding, audio as well as video signal enhancements and interoperability behaviour.

Novel signal processing techniques are necessary to fully exploit the capabilities of specific setups at the acoustic frontend. These signal processing techniques can be roughly separated into two categories:

- Signal enhancement, e.g., noise reduction, echo control, dereverberation, etc.
- Signal transmission, e.g., multi channel coding, error concealment, etc.

In both categories, novel concepts are presented in this thesis along with a novel procedure for the evaluation of the performance of multi channel signal processing schemes. This novel instrumental quality measure is specifically trained for multi channel scenarios and explicitly takes spatial properties of the signals into account.

With respect to the signal enhancement area, a beamforming scheme for an improved recording of signals in adverse environments by utilizing the spatial properties of the acoustic situation is presented. This beamforming scheme was developed within the aforementioned CoVR project, yet the underlying concept is very flexible and allows to employ it in various application scenarios.

The continuous development of array signal processing systems [HL10] throughout the last decades was driven by many applications in the radio frequency domain [HLS93] as well as the acoustic domain [BW01]. A specific form of an array signal processing system in the acoustic domain is the so-called linear microphone array which, due to its physical design, can be integrated easily in many communication systems such as video conferencing clients like the one developed within the CoVR project. A well designed microphone array can exploit the spatial separation between the target and (possibly multiple) interferers. It is thus an efficient way to already achieve a decent *Signal-to-Noise Ratio* (SNR) directly at the acoustic frontend.

An acoustic environment where microphone array systems can be particularly useful can be characterized by a rather low diffuse background noise level and little reverberation. Both of these features are characteristic for a spatial environment where sources (both target and interfering) more closely resemble point sources. Additionally, there has to be a certain spatial separation between the sources. Since this spatial separation is usually given in conferencing scenarios, the use of microphone arrays is especially beneficial in such an environment and a microphone array is an efficient way to simultaneously amplify one target speaker while damping other speakers and background noise.

When designing and parameterizing microphone arrays, the objective is usually to generate a certain reception characteristic. For the far field situation, i.e., at distances from the array that are significantly larger than the physical size of the array setup, there are many known procedures that can be utilized. There are some approaches that are specifically designed for the near field [KAWW96, RG97, RG00, FR11] where the far field designs can only be used to approximately determine the reception characteristic. These approaches however, optimize the reception characteristic on a (semi-)circular arc at one specific distance from the array. A different design was proposed in [ZGET04] which allows to define a target region in the near field and modify the con-

straints for an adaptive beamformer accordingly. No approach is known yet that allows to directly optimize the reception characteristic for an entire area in the near field of the microphone array simultaneously for different distances and angles.

A beamforming algorithm is presented in this thesis for this specific use case: A simulation of the acoustic environment of the microphone array is combined with a numerical optimization procedure to determine filter coefficients for a filter-and-sum beamformer. The integration of the acoustic environment is done based on impulse responses that are simulated or measured between all points of interest in the near field and all microphones in the microphone array.

The numerical optimization procedure requires a meaningful error criterion: This can be found by comparing a target reception characteristic with the current reception characteristic which can be determined via the impulse responses. The performance of the system is evaluated and example configurations are presented. The novel beamforming algorithm is shown to closely approximate the target reception characteristic.

When looking at multi channel signal transmission, there are already quite a few proposals that are useful for certain application scenarios. Here, a system for the transmission of multi channel signals is presented which allows for low algorithmic delay while having advantageous properties that make the transmission very efficient with respect to the data rate. Three aspects in particular are considered:

- Downmixing
- Predictive Coding Techniques
- Noise Shaping in a Multi Channel environment

Even if more and more multi channel equipped devices are recently entering the market (or will do so in the near future), there will always be a rather long transition period where the new multi channel devices coexist with older single channel devices. Due to compatibility reasons, there has to be some way of interaction between the two classes of devices. Downmixing the multiple channels from the newer devices to a mono representation is a simple yet effective way of bridging the gap to the single channel devices.

Hence, an efficient coding scheme for stereo or multi channel signals that includes a downmixing stage is a topic of growing interest. A predictive coding system that exploits the temporal and spatial correlations between the individual channels is advantageous for this scenario as these systems usually have very low algorithmic delay and low computational complexity.

The earliest proposals that can be seen as a simple compressive time-domain predictive encoding scheme for multi channel signals (i.e., not simply by using multiple single channel systems in parallel) date back to the 1950s and 1960s when FM broadcasting was extended to allow for the compatible transmission of stereo signals [Com61] (and later even of quadraphonic signals [DT73]). This system takes correlation between the two channels into account by means of a fixed preprocessing step which generates one sum channel to ensure backwards compatibility to mono receivers.

An overview on more recent developments in time-domain predictive coding of multi channel signals can be found in [Bis07]. A different approach for the coding of stereo signals is to tightly integrate the stereo prediction into the codec. One example is the complex-valued stereo prediction as proposed in [HCD+11] for the *MPEG Unified Speech and Audio Coding* (USAC) approach.

The developments in the area of multi channel coding for applications like storage or streaming (e.g., MPEG Surround [ISO09]) which are not critical with respect to algorithmic delay are summarized in [Her04]. The algorithms from these approaches were recently reconfigured for low-delay operation and tailored for a combination with the AAC-ELD codec as presented in [LVSH11].

One important point for any coding scheme that shall be deployed in the existing telephone network as well as in a high-quality audio conferencing system should be to ensure backwards compatibility. For the single channel case, many known codecs achieve this by being structured in a hierarchical manner (e.g., [ITU06] or [ITU08]) thus allowing to scale the audio bandwidth and the perceived quality according to the available data rate and the capabilities of the hardware that is used. In contrast to that, the aforementioned recent approaches for multi channel coding do not incorporate any possibility for a single channel receiver when combined with codecs that are in use in the telephone network. The presented approach is usable with any mono codec for the main channel ensuring backwards compatibility. An example combination of the proposed coding scheme with the Adaptive Multi-Rate Wideband [ITU03] is presented here.

The evaluation of the perceived quality is an important aspect for the design and development of signal enhancement and transmission systems. There are two possibilities to carry out this quality estimation:

- **Listening tests**
  In a listening test, a group of human listeners has to listen to usually multiple signals and answer one or more questions regarding, e.g., the speech quality. A subsequent statistical analysis of the listeners' responses gives

the final evaluation result. Listening tests are a very flexible evaluation method and they can be tailored exactly to the aspect of the signal processing system that shall be tested. For most of the common tasks, there are standardized procedures (e.g., [ITU96a]) and tools available to conduct listening tests (e.g., [SSGV11]). However, listening tests are very time consuming and hence not suitable for a continuous evaluation during algorithm development.

- **Instrumental quality measures**
  An instrumental quality measure aims at replicating the results of a listening test with a suitably designed calculation rule (e.g., *Perceptual Evaluation of Audio Quality* (PEAQ) [ITU01a]). The drawback of this approach is a loss of flexibility: If there is no instrumental measure available for the aspect that shall be evaluated, listening tests will be the only possibility. The advantage of this approach on the other hand is obvious: The instrumental measure allows to easily evaluate the perceived quality of a signal enhancement or transmission system during algorithm development.

Hence, the instrumental assessment of the perceived quality is a topic that has been receiving continuous interest. An overview on instrumental quality assessment in general can be found in [RBK+06, Côl1]. So far, only basic approaches for the evaluation of multi channel signals in particular were considered in [GZR06, ZRKB05].

A novel, more advanced concept for the quality evaluation of multi channel signal processing systems is presented here. Its foundation is a coincidence-based binaural hearing model which consists of a physiologically motivated signal processing step and a subsequent cognitive model. Particular care has been taken to ensure robustness and the model is shown to be capable of correctly detecting and tracking sources even in adverse acoustic environments and for multiple concurrent speakers. It is also capable of blindly determining the number of sources that are currently active which can make it an interesting enhancement to source separation algorithms which often rely on this knowledge. Spatial parameters are derived from this model which are then combined with the output of PEAQ, a known algorithm for the evaluation of audio quality, to get a joint measure for audio quality and spatial fidelity.

Without the additional parts for quality evaluation, the hearing model alone is of interest in all areas that need to consider the capabilities of the human hearing system. Of the various parameters that can be derived from the hearing model, the five most important spatial parameters for the perceived quality are determined based on listening tests and subsequently utilized for a quality

measure. This measure is derived with a methodology that is very similar to the basis for the PEAQ measure: the aforementioned most important spatial parameters are fed into a *Neural Network* (NN) together with the result of PEAQ to get the overall result. The inclusion of spatial information into the instrumental quality measurement leads, in contrast to PEAQ, to a consistently high correlation between the instrumental measure and a listening test for stereo signals.

1.1 Structure of the Thesis

In Chapter 2, some fundamentals of multi channel signal processing are presented. This starts with some remarks on the acoustic environment which lead to a signal model (and notation) that is used throughout the thesis. Since a major part of the thesis deals with predictive coding systems, the historical context and important aspects of these systems are described as well. A novel way of determining the filter taps in a multi channel predictive system is devised and its performance is analyzed. This novel way is based on an alternative interpretation of linear prediction as a model building procedure.

The two-stage system which forms the core of this thesis is described and analyzed in Chapters 3 and 4. The outer stage of this system, i.e., the first stage on the transmitting side and the second stage on the receving side, is analyzed in Chapter 3. Different use cases for this stage are considered on the transmitting and receiving side. One of these use cases is for the transmission of multi channel signals: The transmitting side of the outer stage then carries out downmixing to match the number of input channels to the number of channels that are transmitted. The task of the receiving side is then to match the number of transmitted channels to the reproduction setup, i.e., the number of loudspeakers.

The outer stage can also fulfil other tasks. One important possibility on the transmitting side is beamforming. Especially if the number of microphones is large in comparison to the number of transmitted channels, beamforming is a good possibility for signal enhancement directly at the acoustic frontend. A numerical optimization scheme that is based on knowledge about the acoustic environment is introduced and evaluated. The novel scheme allows to define a reception characteristic in the near field of the microphone array which is then approximated by optimizing the coefficients of the beamforming system. On the receiving side, a postfiltering system targeted at improving the speech intelligibility is presented which takes its inspiration from knowledge about

room acoustics. Different rooms and their properties are analyzed and an optimized system is derived that is shown to increase speech intelligibility.

The inner stage of the two-stage system, i.e., the second stage on the transmitting side and the first stage on the receiving side, as described in Chapter 4 has one major application: redundancy removal for an efficient transmission. The techniques that are used in this stage stem from the realm of predictive coding. The fundamental system has an advantageous property for practical use: There is a very simple and seamless possibility to connect a multi channel client to a single channel client since the multi channel coding scheme uses a single channel downmix as one of its core elements.

The multi channel coding scheme is thoroughly analyzed and choices for its parameters are derived. A novel concept for incorporating noise shaping into multi channel predictive coding schemes is presented and different variants are explained. A combination of single channel open loop predictors and a cross channel noise feedback is shown to give the best results in an application example.

The application example utilizes a novel multi channel quality estimation system which is presented in detail in Chapter 5. The system is an extension to the well-known PEAQ measure which performs well for single channel signals but fails to correctly incorporate spatial properties into its quality estimation. To overcome this issue, a specifically tailored binaural hearing model is developed which provides numerous spatial parameters for the novel instrumental measure. The final quality score is a combination of the result of PEAQ with the spatial parameters. On the basis of a listening test, this combination is shown to clearly outperform PEAQ alone.

1.2 System Overview

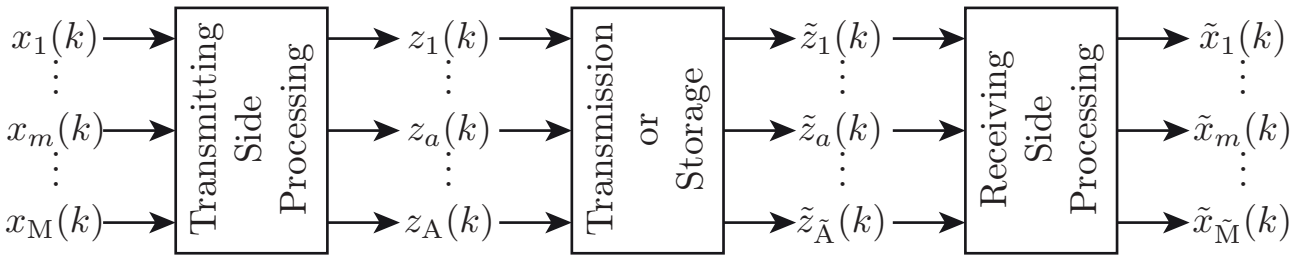


Figure 1.1: Transmitting side of the two-stage system

In parts of Chapter 2 and especially in Chapters 3 and 4, the transmitting side and the receiving side processing are described in more detail. The two-stage structure of the transmitting side of the system is depicted in Figure 1.2.

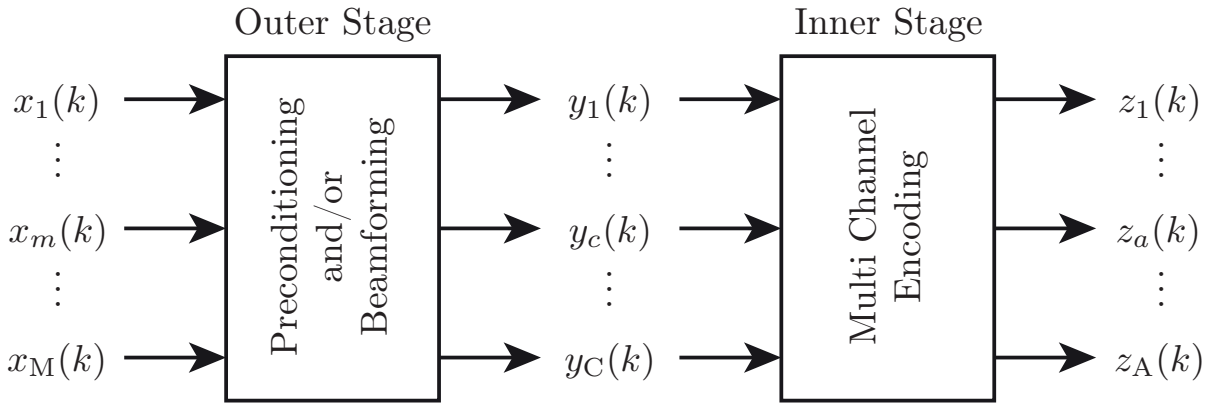


Figure 1.2: Transmitting side of the two-stage system

The M input (or microphone) signals are labeled $x_m(k)$, the C transmitting side intermediate signals between the two stages are labeled $y_c(k)$ and the A transmission signals are denoted $z_a(k)$. The receiving side of the signal processing and transmission system is depicted in Figure 1.3. On this end, the $\tilde{\mathrm{A}}$ reception signals are labeled $\tilde{z}_m(k)$, the $\tilde{\mathrm{C}}$ receiving side intermediate signals are denoted $\tilde{y}_c(k)$ and the $\tilde{\mathrm{M}}$ output (or loudspeaker) signals are labeled $\tilde{x}_m(k)$. Note that it not necessarily the target to perfectly reconstruct the input signals at the output of the system since some of the proposed systems (especially the beamforming algorithm) for the outer stage on the transmitting side will not aim at conserving these signals but at exploiting certain properties of the signals.

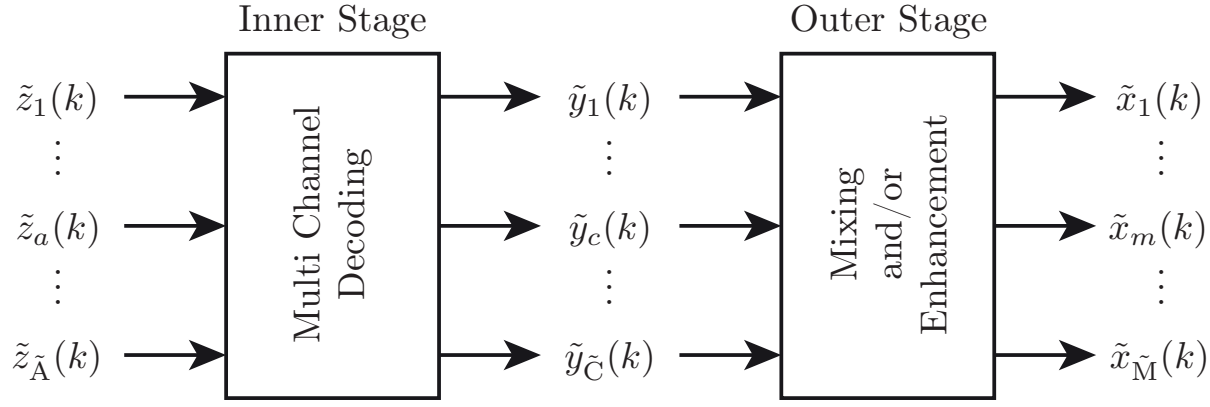


Figure 1.3: Receiving side of the two-stage system

The outer stage will be treated in Chapter 3 and used to either perform spatial filtering or signal preconditioning depending on the application, the inversion of this stage is either used to increase the speech intelligibility of the received signal or to generate the output signals via mixing. The task of the inner stage is to decrease the necessary data rate for transmission of the signals – a

hierarchical multi channel predictive coding system for this will be presented and analyzed in Section 4.

Parts of the results of this thesis have been pre-published in the following references: [SKV09, SJSV10, SSGV11, SBV12, SV12b, SHWV12, SBV13, JSV09, JSEV10, JSK+10, GSV11, HSV+12, HSWV13, SHS+13, BFS+13]. These references are marked by an underlined label, i.e., [____], throughout the thesis.

2

Multi Channel Signal Processing – Fundamentals and New Developments

In this thesis, different aspects of multi channel signal processing in a communication scenario are considered. Starting at the acoustic front end, beamforming or channel mixing are used, depending on the application scenario, to exploit or conserve the spatial information that is present within the microphone signals. The output signals of this first stage are then transmitted by means of a predictive coding scheme. The evaluation of the perceived quality of the output signals at the receiving end is the final part of the presented methodologies.

For all these parts, a common signal model and notation as well as some fundamentals of signal processing and acoustics are introduced in this chapter. Some more detailed evaluations on specific aspects of the system form the rest of the chapter.

As a main aspect of the thesis is the large area of linear predictive systems, these methods receive special treatment in this chapter.

An alternative formulation of the prediction process is proposed which paves the way for a novel way of determining the filter coefficients (and filter delays) of generic linear predictive systems and a new system is presented that allows to perform joint multi channel linear predictive coding with adaptive distribution of the available data rate between coefficients for intra channel and inter channel prediction.

Firstly, the signal model that is used throughout the following chapters is introduced in Section 2.1 and properties of acoustic environments and multi

channel signals are presented in Sections 2.2 and 2.3. The larger part on linear prediction techniques begins with a short review of linear predictive coding in general in Section 2.4 and the limits thereof in Section 2.5. A multi channel prediction system based on an alternative formulation of linear prediction is described and evaluated in Section 2.6.

2.1 Discrete-Time Signal Model

The plain recording of multi channel signals (i.e., not for artificially produced multichannel material, e.g., in music productions) can be represented in the free field by a simple signal model: An arbitrary sound wave travelling from a source through a space with M microphones as shown in Figure 2.1 which leads to a distinct transmission characteristic to each individual microphone.

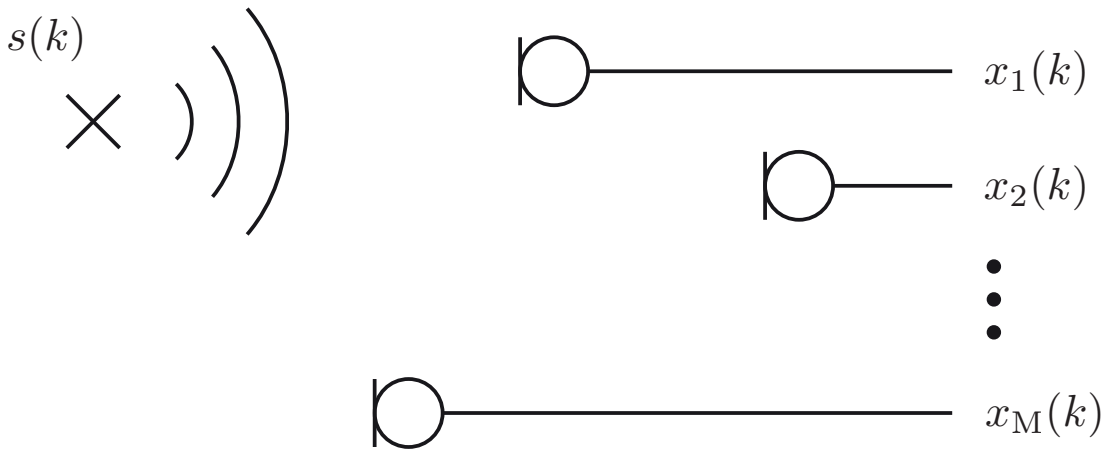


Figure 2.1: Simple free field signal model

Assuming the discrete-time signal $s(k)$ (sampled at time instants $k \cdot T$ with the sampling interval T and the sampling frequency $f_s = \frac{1}{T}$) is emitted by the source depicted in Figure 2.1, every microphone m will pick up a delayed and damped version of $s(k)$:

$$\begin{aligned} x_1(k) &= \alpha_1 \cdot s(k - \tau_1) \\ x_2(k) &= \alpha_2 \cdot s(k - \tau_2) \\ &\vdots \\ x_m(k) &= \alpha_m \cdot s(k - \tau_m) \\ &\vdots \\ x_{\mathrm{M}}(k) &= \alpha_{\mathrm{M}} \cdot s(k - \tau_{\mathrm{M}}) \end{aligned} \tag{2.1}$$

Both the damping α and the delay τ (expressed in the number of samples) are directly related to the distance d_m between the source and the respective microphone by:

$$\alpha_m = \frac{1}{d_m} \tag{2.2}$$

and

$$\tau_m = \frac{d_m}{c_0 \cdot T} \tag{2.3}$$

with the speed of sound c_0. In the following, the sampling frequency is assumed to be high enough (or the signals suitably interpolated) such that the delay equals an integer number of samples, i.e., $\tau_m \in \mathbb{Z}$[1].

It can be seen that there is also a very simple relation between two microphone signals (e.g., $x_1(k)$ and $x_2(k)$). Depending on the distances (d_1 and d_2) of the two microphones to the source and the related dampings (α_1 and α_2) and delays (τ_1 and τ_2), a (possibly non-causal) filter can be derived that allows to calculate $x_2(k)$ from $x_1(k)$:

$$x_2(k) = \frac{\alpha_2}{\alpha_1} \cdot \delta(k - (\tau_2 - \tau_1)) * x_1(k) \tag{2.4}$$

with $*$ indicating convolution and $\delta(\cdot)$ as the Kronecker delta.

However, while this scenario allows for a first analysis of the behaviour of simple multi channel recording setups, it is only of low importance for real-world signals since there is almost no environment where the underlying free-field assumption of Figure 2.1 is fulfilled. Hence, the more general setup depicted in Figure 2.2 will be considered in the following. The same setup as in Figure 2.1 is put into an enclosure, leading to a significant influence of room reflections.

The simple delay and damping from source to microphone is thereby replaced by a *Finite Impulse Response* (FIR) filter that characterizes the transmission from source to microphone within the room, the so-called *Room Impulse Response* (RIR) $h_m(k)$. The microphone signals $x_m(k)$ can then be expressed by the convolution of the source signal $s(k)$ with the respective RIR (which is

[1] Fractional delays [VL93, MKK94] would also be possible but are not treated here – while they do not lead to additional insights, they make the analyses less accessible

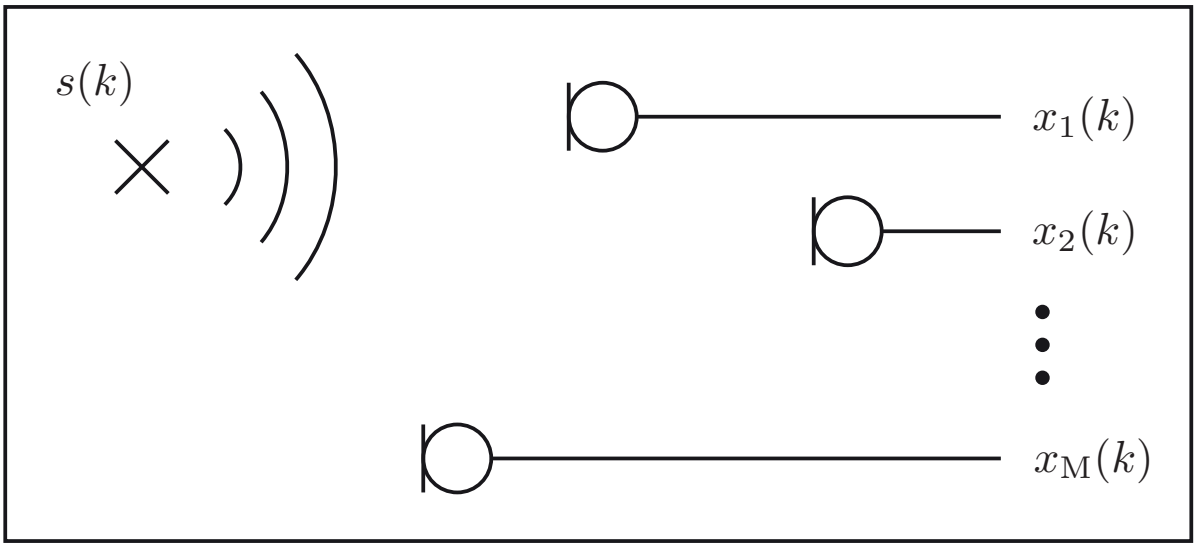


Figure 2.2: Simple reverberant signal model

different for each microphone):

$$\begin{aligned} x_1(k) &= h_1(k) * s(k) \\ x_2(k) &= h_2(k) * s(k) \\ &\vdots \\ x_m(k) &= h_m(k) * s(k) \\ &\vdots \\ x_{\mathrm{M}}(k) &= h_{\mathrm{M}}(k) * s(k) \end{aligned} \tag{2.5}$$

The aforementioned simple delay and damping can also be expressed as a special case of the RIR by

$$h_m(k) = \begin{cases} \alpha_m & k = \tau_m \\ 0 & \text{else.} \end{cases} \tag{2.6}$$

With the more general filter $h_m(k)$, it gets significantly more complicated to calculate microphone signal $x_2(k)$ from $x_1(k)$. The necessary deconvolution is a topic beyond the scope of this thesis, an overview on this topic can be found in [Buc10].

The RIR is a complete representation of the linear part[2] of the acoustic characteristics of the transfer path between the source and the microphone. The

[2]Note that the influence of non-linearities is negligible for most acoustic environments in the area of speech and audio transmission.

equivalent quantity in the discrete frequency domain is the *Transfer Function* (TF) $H_m(\mu)$. With the frequency index μ and the transform length N, the TF can be calculated according to

$$H_m(\mu) = \mathcal{F}\{h_m(k)\} = \sum_{i=0}^{N-1} h_m(k) \cdot \mathrm{e}^{\frac{j2\pi\mu i}{N}} \tag{2.7}$$

2.2 Generation and Measurement of Room Impulse Responses

It is very important for the design and evaluation of speech and audio signal processing systems to take the properties of the acoustic environment into account and to verify the usability of, e.g., signal enhancement algorithms not only for ideal, theoretic acoustic situations. Even for telephony applications, the effect of room reverberation on the transmitting side can not be neglected as shown in [JSK+10] based on objective speech quality measures as well as a listening test.

There are different ways to generate RIRs artificially based on parameters of the acoustic environment, e.g., in a very quick and simple manner by the source-image method [AB79] or in a much more precise and sophisticated way in auralization systems for virtual acoustic environments [Sch12]. Alternatively, some datasets that were measured in real rooms are available to the scientific community that contain RIRs for vastly different acoustic environments and for varying source and microphone setups.

One of these data sets is the so-called *Multichannel Acoustic Reverberation Database at York* (MARDY) which was introduced and described in [WGH+06]. All the measurements in this database were taken in a specific measurement room which is equipped with interchangeable panels that can be moved to change the acoustic properties of the room. The authors measured the RIRs at different source-microphone distances with eight microphones at inter element spacings of 0.05 m.

An extensive database of measured *Binaural Room Impulse Responses* (BRIRs), the so-called *Aachen Impulse Response* (AIR) database, was presented in [JSV09]. It consists of recordings in a low-reverberant studio booth, an office room, a meeting room and a lecture room, all measured with and without a dummy head. The database was subsequently extended (e.g., in [JSK+10]) with additional measurements of other acoustic environments, especially some strongly reverberant rooms were added to the portfolio.

The measurement of room impulse responses in real environments has, in comparison to the simulation approaches, the decisive advantage that the acoustic properties of the measurement room can be accurately reproduced. They can be obtained very effectively with pseudorandom sequences (e.g. *Maximum Length Sequences* (MLSs), cf. [RV87], or *Perfect Sequences* (PSEQs) cf. [MHA08, Chapter 7]).

An important parameter of acoustic environments is the reverberation time T_{60} which is defined as the time that it takes for the sound pressure level to decrease by 60 dB. The reverberation time depends on the size of the room and the shape and the reflective properties of the surfaces in the room. Typical reverberation times are in the range of a few hundred milliseconds for small and medium rooms (e.g., office rooms, living rooms) up to about one second for larger rooms (e.g., lecture halls, stairways) while they can be even longer for very large rooms with many hard, reflective surfaces. Large cathedrals can thus have reverberation times of more than ten seconds.

Not only does the reverberation time give important information about the acoustic environment but it can also be used to determine certain system dimensions since it quantifies the number of significant coefficients of the RIR. The postfiltering approach that will be presented in Section 3.4 is motivated by the impact of reverberation on speech intelligibility and could be parameterized by the length of the RIR, i.e., the reverberation time.

2.3 Properties of Multi-Channel Signals

A major part of this thesis, Chapter 4, focusses on the the design of transmission systems for multi channel signals. These systems achieve a data rate reduction by exploiting certain properties of the input signals. This section reviews some of the general and specific properties of classes of multi channel signals that will be used later on for the evaluation of the presented transmission systems.

There are two fundamental classes of multi channel signals that any multi channel transmission system should be able to encode and decode efficiently both with respect to the necessary data rate as well as to the computational complexity:

- **Natural multi channel Signals**
  These signals consist of multiple recordings of any acoustic event. A model can be utilized that can be derived from the acoustic properties of the source-microphone setup that was used (cf. Section 2.1).

- **Artificial multi channel Signals**
  These signals are generic multi channel signal where no specific production or recording model can be assumed. They can be, e.g., the result of post-processing of audio recordings in the music production process.

Even though the former can be viewed as a special case of the latter, it is such a common and important special case that it shall definitely be analyzed separately.

2.3.1 Simple Recording Model

The first class of multi channel signals, the recording of multi channel signals in realistic acoustic environments, can be directly derived from the signal model in Section 2.1 with the only decisive change that there may also be a convolutive mixture, i.e., every microphone signal $x_m(k)$ contains multiple concurrent sources $s_1(k) \dots s_{\mathrm{N}}(k)$ that are filtered according to the different acoustic paths.

$$\begin{aligned}
x_1(k) &= h_{1,1}(k) * s_1(k) + \cdots + h_{n,1}(k) * s_n(k) + \cdots + h_{\mathrm{N},1}(k) * s_{\mathrm{N}}(k) \\
x_2(k) &= h_{1,2}(k) * s_1(k) + \cdots + h_{n,2}(k) * s_n(k) + \cdots + h_{\mathrm{N},2}(k) * s_{\mathrm{N}}(k) \\
&\vdots \\
x_m(k) &= h_{1,m}(k) * s_1(k) + \cdots + h_{n,m}(k) * s_n(k) + \cdots + h_{\mathrm{N},m}(k) * s_{\mathrm{N}}(k) \\
&\vdots \\
x_{\mathrm{M}}(k) &= h_{1,\mathrm{M}}(k) * s_1(k) + \cdots + h_{n,\mathrm{M}}(k) * s_n(k) + \cdots + h_{\mathrm{N},\mathrm{M}}(k) * s_{\mathrm{N}}(k)
\end{aligned} \tag{2.8}$$

As before, the filter $h_{n,m}(k)$ represents the RIR. The two indices are indicating that the RIR from signal source n to the microphone m is meant. It can be seen that all N source signals are present in all M microphone signals.

2.3.2 Artificial Multi Channel Signals

In contrast to the multi channel signals that were recorded in a real natural environment, artificial mixtures are significantly more difficult to model. The signals $x_1(k) \dots x_{\mathrm{M}}(k)$ (which should here not be understood as microphone signals but in a more general way as the output of an arbitrary signal generator) have no fixed relation at all, e.g., they can be convolutive mixtures as in

Section 2.3.1, they can be completely independent and they can be related in some unnatural way (e.g., $x_1(k) = -x_2(k)$).

However, in many possible applications for multi channel speech and audio, artificial mixing is not utilized to create unnatural acoustic scenarios but, e.g., to render the participants of a conference to distinct positions. Hence, most artificial mixtures do at least resemble natural signals. Nevertheless, a comprehensive analysis of the properties requires to look not only at convolutive mixtures but at the relations between the channels on a more generic basis.

2.3.3 Coherence between the Channels

It is intuitively clear that there has to be a relation between the channels to facilitate any gain from joint encoding later on in Chapter 4. The generic measure to quantify the relation between the input channels $x_{m_1}(k)$ and $x_{m_2}(k)$ is their cross-correlation $\varphi_{m_1 m_2}(\lambda)$.

$$\varphi_{m_1 m_2}(\lambda) = \mathrm{E}\left\{x_{m_1}(k) \cdot x_{m_2}(k-\lambda)\right\} \tag{2.9}$$

In the frequency domain, a normalized measure for the correlation between the channels $x_{m_1}(k)$ and $x_{m_2}(k)$ is the coherence, which is often used in the *Magnitude Squared Coherence* (MSC) representation $\Gamma_{m_1 m_2}(\mu)$ [VM06].

$$\Gamma_{m_1 m_2}(\mu) = \frac{\left|\Phi_{m_1 m_2}(\mu)\right|^2}{\Phi_{m_1 m_1}(\mu) \cdot \Phi_{m_2 m_2}(\mu)} \tag{2.10}$$

with the *Power Spectral Density* (PSD) $\Phi(\mu)$ as the frequency transform of the correlation $\varphi(\lambda)$.

The MSC only allows values between 0 and 1 where a value of $\Gamma_{m_1 m_2}(\mu) = 0$ represents completely independent signals while a value of $\Gamma_{m_1 m_2}(\mu) = 1$ is reached for two signals as soon as one signal is a linearly filtered version of the other:

$$x_{m_1}(k) = h(k) * x_{m_2}(k) \tag{2.11}$$

Large gains from a joint encoding can be expected if the coherence values become large. For any signal that is recorded in a real environment as described in Section 2.3.1, the coherence should be $\Gamma_{m_1 m_2}(\mu) \approx 1$ as long as the transmission scenario is linear.

However, this only holds for infinite signal lengths of $x_{m_1}(k)$ and $x_{m_2}(k)$. In most digital transmission systems, the processing is done on small blocks of data, so-called frames, which have a length of only a few milliseconds. In these frames, a commonly applied algorithm for the estimation of the coherence is the Welch-periodogram approach [Wel67]. Like any estimation algorithm, this approach has some peculiar properties that have to be considered in order to be able to correctly interpret the resulting values.

There is, for example, a strong dependency between the reverberation time T_{60} of the recording room, the length of the frames within the signal processing algorithm and the estimated short-time coherence in that frame. A closer look at this relation can be found in [SV10] where the coherence estimation is carried out with different frame lengths. The target of that evaluation being the correct parametrization of a dereverberation algorithm. The brief results of the evaluation are that $\Gamma_{m_1 m_2}(\mu) \approx 0$ results for short frame lengths while $\Gamma_{m_1 m_2}(\mu) \approx 1$ results for long frames and the shape of the *Coherence Estimate Function* (CEF) depends on the acoustic environment.

For the performance analysis later on in Section 2.5, it is worth noting that the performance bound of the inter channel prediction scales with the coherence that is present and hence it strongly depends on the frame length and the look-back and look-ahead of the coding scheme.

2.4 Linear Prediction

The so-called linear predictive analysis is probably one of the best known concepts for the parametric representation of the spectral envelope of many natural signals in general and speech signals in particular [Yul27, Mak75]. An overview of this can be found in, e.g., [VM06, JN84], only a short recapitulation of the most important principles for this work will be given here.

Classical linear prediction in the single channel case is based on the source-filter model of human speech production which consists of a source (lungs and vocal folds) emitting either pulse trains or noise sequences and a filter (vocal tract) that shapes the spectrum of the signal. Both parts are generally time-varying but can be assumed to be short-term stationary for the following analyses.

2.4.1 Single Channel Linear Prediction

In the most widely used version of single channel *Linear Prediction* (LP), temporal correlation between successive signal samples is removed by means of an

FIR filter that is updated frequently since, as mentioned, speech and audio signals can only be assumed to be stationary for short periods of time. Common choices for these update intervals are in the range of 16 to 32 milliseconds.

A simple forward predictor is depicted in Figure 2.3, where the current sample $y(k)$ is predicted from the past L samples by means of the prediction filter $\mathbf{H}$.

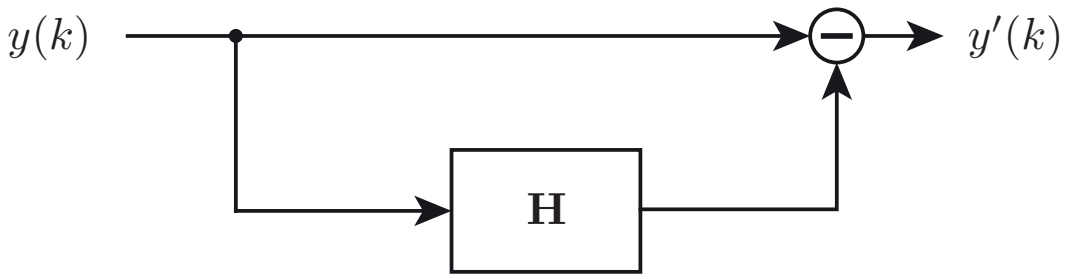


Figure 2.3: Single channel linear predictor

$$\mathbf{H} = \Big(\, h(1) \; h(2) \; \ldots \; h(L) \, \Big)^{\mathrm{T}} \tag{2.12}$$

This results in the prediction error $y'(k)$:

$$y'(k) = y(k) - \sum_{\lambda=1}^{L} h(\lambda) \cdot y(k-\lambda) \tag{2.13}$$

Long Term Prediction

An additional step that is especially suitable for speech transmission systems is the so-called *Long Term Prediction* (LTP) that is depicted in Figure 2.4. The rationale behind this additional predictor stems from the model of human speech production: If a voiced sound is produced by any human speaker, the excitation signal that is produced by the lungs and vocal folds will approximate a periodic pulse train. The prediction error $y'(k)$ after the first prediction stage strongly resembles the excitation signal since all influences of the vocal tract can ideally be removed by the first prediction filter $\mathbf{H}$. Hence, there will be no more significant short term correlation within the signal but a non-negligible correlation at the distance between the pulses in the excitation signal for voiced sounds.

The mostly used LTP filter is conceptually different from the regular LP filter since it has got two different degrees of freedom:

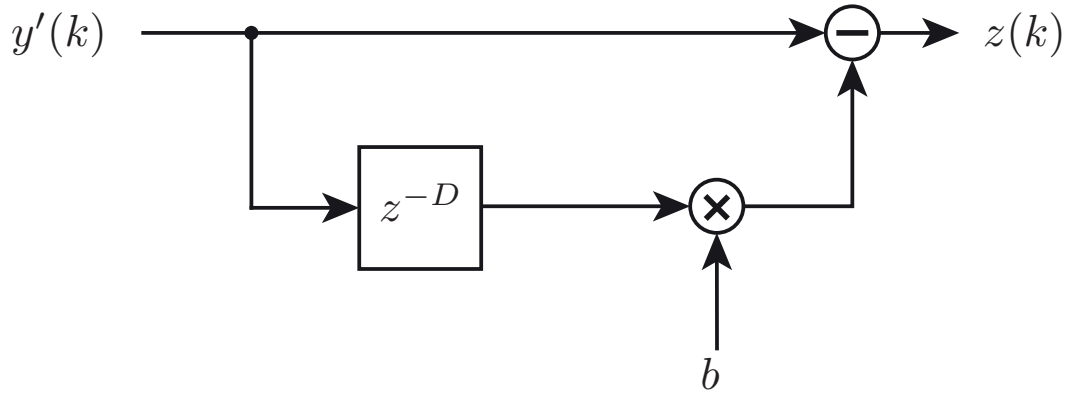


Figure 2.4: Single channel long term predictor

- Filter coefficient b
- Filter delay D

With this additional step, the resulting prediction error $z(k)$ of this stage can be written as:

$$z(k) = y'(k) - b \cdot y'(k - D) \tag{2.14}$$

Inserting Equation 2.13 leads to this expression of the overall prediction error:

$$z(k) = y(k) - \left(\sum_{\lambda=1}^{L} h(\lambda) \cdot y(k - \lambda) + y(k - D) - \sum_{\lambda=1}^{L} h(\lambda) \cdot y(k - (D + \lambda)) \right) \tag{2.15}$$

Similar equations can be derived for every possible setup and combination of predictive filtering systems. In the following chapters, similar structures will be analyzed with respect to their properties regarding their encoding performance.

2.4.2 Determination of Optimal Filter Coefficients

Predictive transmission systems usually have the target to decrease the energy of the frame of signal samples $z_a(k)$ to be transmitted. In single channel systems, this prediction is done by exploiting temporal relations within the signal to predict the current sample from the recent past. In multi channel systems, both the temporal relations within each channel can be used by means of so-called *intra-channel* prediction as well as the (possibly spatial) relations between the channels by means of so-called *inter-channel* prediction.

For both types of prediction, filter coefficients have to be determined that minimize the *Minimum Mean Square Error* (MMSE) of the spectrally shaped prediction error signal. The parameters of the prediction system can be determined by calculating the error measure and differentiating it with respect to the parameter that shall be determined.

The MMSE is clearly the most prominent criterion for prediction tasks, an overview on the derivation of optimum filter coefficients in single channel applications with this criterion can be found in, e.g., [VM06]. It is also the criterion that is used throughout this thesis to determine system parameters.

2.4.3 Stability at the decoder

In all signal transmission systems, the decoding stage has to invert the processes that took place in the encoding stage. In single channel linear prediction systems this means that since the encoder contains an FIR filter, the decoder has to contain an *Infinite Impulse Response* (IIR) filter for the reconstruction. Depending on the way that the filter coefficients are determined and quantized, respectively, this can result in a filter at the decoding side that is not stable. There are techniques to ensure a stable decoder by modification of the filter structure (e.g., into a lattice structure) resp. a transformation of the the filter coefficients into reflection coefficients (with appropriate limiting of their absolute value), *Line Spectral Pairs* (LSPs) or *Line Spectral Frequencies* (LSF) [SJ84, KR86]. In addition, the LSP or LSF representation also exhibits nice properties for estimating missing parameters after transmission, a scheme for this based on *Gaussian Mixture Models* (GMMs) was presented in [MHW01].

In contrast to that, the necessary inversion does not always necessitate an inverted filter structure in multi channel prediction systems depending on the specific system design. (E.g., the system that is described in Chapter 4 partly uses predictive filters between the channels in a setup that does not require the inversion of the filters themselves but only of the sign of the filter coefficients.)

2.5 Performance Limit

A performance limit for multi channel prediction systems can not be directly derived from a combination of the known results for intra and inter channel prediction since the filter coefficients in a combination of those cases are not optimized jointly but sequentially. A comparison of the results for different optimization strategies is carried out in this section.

Looking at the intra channel case alone, the so-called spectral flatness [MW72, MA76] provides a suitable measure to quantify the achievable prediction gain depending on the properties of the input signal. This is due to the fact that the optimum output of a predictive system that only does intra channel prediction is a spectrally white signal. The spectral flatness can be calculated based on the power spectrum of the signal by dividing the geometric mean of this spectrum by the arithmetic mean of this spectrum.

In the inter channel case for two signals, the MSC (cf. Section 2.3.3) between the input signals can be utilized to give an indication of the achievable performance of a prediction system that filters one input signal to minimize the difference to the other input signal. A closer look at this in the area of dual channel noise reduction can be found in [VM06] the results of which directly relate to prediction systems as well.

A block diagram of the multi channel prediction system that is the basis for the following analyses is depicted in Figure 2.5. Two filters $\mathbf{h}_{\text{intra}}$ and $\mathbf{h}_{\text{inter}}$ are utilized with L_{intra} and L_{inter} coefficients, respectively. These coefficients can be derived sequentially or jointly by means of two MMSE optimizations or one MMSE optimization, respectively. The analyses are carried out here for the two channel case but they readily extend to more channels by introducing an additional sum over the other input channels in the calculation of the prediction errors in Eqs. 2.18 and 2.20.

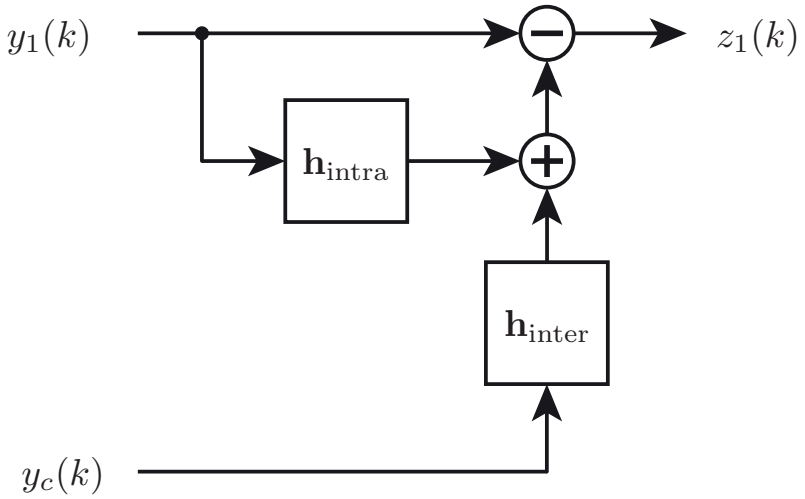
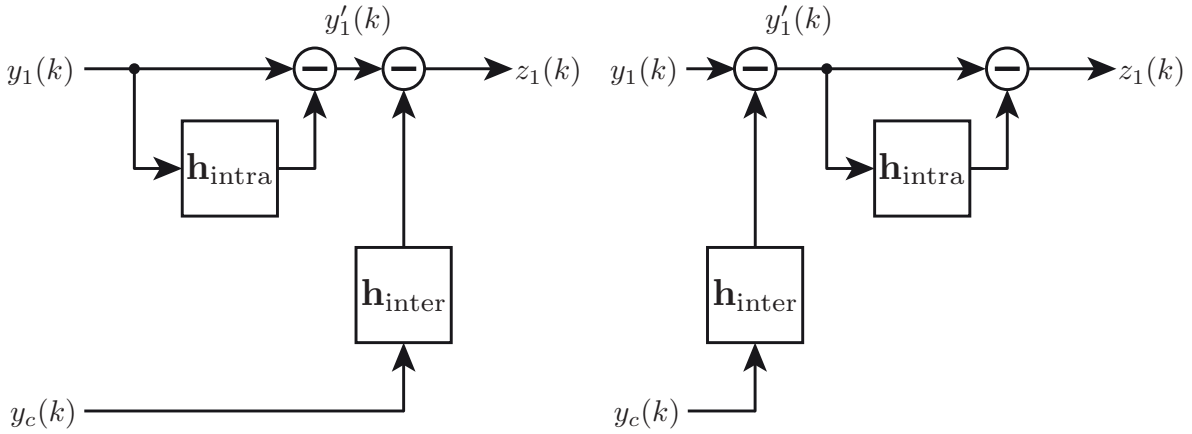


Figure 2.5: Joint intra and inter channel prediction system

For the sequential setup, there are two possible setups:

- Doing the intra channel prediction first (see Figure 2.6a)
- Doing the inter channel prediction first (see Figure 2.6b)



(a) Intra channel prediction first (b) Inter channel prediction first

Figure 2.6: Different possible setups for the sequential optimization

The first step in both sequential variants is the minimization of the expectation of the intermediate prediction error signal $y_1'(k)$. The expectation here is equivalent to the short-term energy of the prediction error signal since the processing is done in a framewise manner such that the signals are inherently limited in time.

$$\mathrm{E}\left\{y_1'(k)^2\right\} \doteq \sum_{i=0}^{N} \left(y_1'(k-i)\right)^2 \rightarrow \min \tag{2.16}$$

After that, the subsequent step – minimizing the short-term energy of the prediction error signal $z_1(k)$ – is identical to the only optimization criterion that is necessary for the joint variant:

$$\mathrm{E}\left\{z_1(k)^2\right\} \rightarrow \min \tag{2.17}$$

Note that the current value of one input signal $y_1(k)$ shall be predicted and can hence not be used in the prediction while the current value of the other input signals $y_2(k) \ldots y_{\mathrm{M}}(k)$ can be used, this will lead to different ranges for λ in the sums in the prediction error signals.

In the joint case and when the intra prediction is carried out first, the prediction error $z_1(k)$ is calculated by

$$z_1(k) = y_1(k) - \sum_{\lambda=1}^{L_{\text{intra}}} h_{\text{intra}}(\lambda) \cdot y_1(k-\lambda) - \sum_{\lambda=0}^{L_{\text{inter}}-1} h_{\text{inter}}(\lambda) \cdot y_c(k-\lambda). \tag{2.18}$$

If the inter channel prediction is carried out first, the prediction error is given

as

$$z_1(k) = y_1'(k) - \sum_{\lambda=1}^{L_{\text{intra}}} h_{\text{intra}}(\lambda) \cdot y_1'(k-\lambda) \tag{2.19}$$

with

$$y_1'(k) = y_1(k) - \sum_{\lambda=0}^{L_{\text{inter}}-1} h_{\text{inter}}(\lambda) \cdot y_c(k-\lambda). \tag{2.20}$$

The filter coefficients for $\mathbf{h}_{\text{intra}}$ and $\mathbf{h}_{\text{inter}}$ are determined differently depending on the way that the optimization is carried out. In the following, both the two different sequential and the joint optimization of the filter coefficients are analyzed and the differences are discussed.

2.5.1 Sequential Optimization

The sequential optimization is done in two steps which can be analyzed independently. One step is classical single channel linear prediction on $y_1(k)$ if the intra channel prediction is carried out first or on $y_1'(k)$, respectively, if the inter channel prediction is carried out first. Optimizing $y_1'(k)$ and $z_1(k)$, respectively, in the MMSE sense leads to the well-known Yule-Walker equations. The equations can be denoted either (if the intra channel prediction is carried out first, cf. Figure 2.6a) as

$$\begin{pmatrix} \varphi_{y_1y_1}(0) & \varphi_{y_1y_1}(1) & \dots & \varphi_{y_1y_1}(L_{\text{intra}}-1) \\ \varphi_{y_1y_1}(1) & \varphi_{y_1y_1}(0) & \dots & \varphi_{y_1y_1}(L_{\text{intra}}-2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{y_1y_1}(L_{\text{intra}}-1) & \varphi_{y_1y_1}(L_{\text{intra}}-2) & \dots & \varphi_{y_1y_1}(0) \end{pmatrix} \cdot \dots \cdot \begin{pmatrix} h_{\text{intra}}(1) \\ h_{\text{intra}}(2) \\ \vdots \\ h_{\text{intra}}(L_{\text{intra}}) \end{pmatrix} = \begin{pmatrix} \varphi_{y_1y_1}(1) \\ \varphi_{y_1y_1}(2) \\ \vdots \\ \varphi_{y_1y_1}(L_{\text{intra}}) \end{pmatrix} \tag{2.21}$$

or (if the inter channel prediction is carried out first, cf. Figure 2.6b) as

$$\begin{pmatrix} \varphi_{y_1'y_1'}(0) & \varphi_{y_1'y_1'}(1) & \dots & \varphi_{y_1'y_1'}(L_{\text{intra}}-1) \\ \varphi_{y_1'y_1'}(1) & \varphi_{y_1'y_1'}(0) & \dots & \varphi_{y_1'y_1'}(L_{\text{intra}}-2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{y_1'y_1'}(L_{\text{intra}}-1) & \varphi_{y_1'y_1'}(L_{\text{intra}}-2) & \dots & \varphi_{y_1'y_1'}(0) \end{pmatrix} \cdot \dots \cdot \begin{pmatrix} h_{\text{intra}}(1) \\ h_{\text{intra}}(2) \\ \vdots \\ h_{\text{intra}}(L_{\text{intra}}) \end{pmatrix} = \begin{pmatrix} \varphi_{y_1'y_1'}(1) \\ \varphi_{y_1'y_1'}(2) \\ \vdots \\ \varphi_{y_1'y_1'}(L_{\text{intra}}) \end{pmatrix} . \tag{2.22}$$

The other part is a prediction either between a prediction error $y_1'(k)$ of the intra channel prediction and an input signal $y_c(k)$ or between the two input signals $y_1(k)$ and $y_c(k)$. This leads to structures very similar to the aforementioned Yule-Walker equations with the decisive difference that there are not only auto-correlation values but also cross-correlation values within the equations. If the intra channel prediction is carried out first, these equations are

$$\begin{pmatrix} \varphi_{y_cy_c}(0) & \varphi_{y_cy_c}(1) & \dots & \varphi_{y_cy_c}(L_{\text{inter}}-1) \\ \varphi_{y_cy_c}(1) & \varphi_{y_cy_c}(0) & \dots & \varphi_{y_cy_c}(L_{\text{inter}}-2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{y_cy_c}(L_{\text{inter}}-1) & \varphi_{y_cy_c}(L_{\text{inter}}-2) & \dots & \varphi_{y_cy_c}(0) \end{pmatrix} \cdot \dots \cdot \begin{pmatrix} h_{\text{inter}}(0) \\ h_{\text{inter}}(1) \\ \vdots \\ h_{\text{inter}}(L_{\text{inter}}-1) \end{pmatrix} = \begin{pmatrix} \varphi_{y_1'y_c}(0) \\ \varphi_{y_1'y_c}(1) \\ \vdots \\ \varphi_{y_1'y_c}(L_{\text{inter}}-1) \end{pmatrix} . \tag{2.23}$$

Doing the inter channel prediction first leads to

$$\begin{pmatrix} \varphi_{y_cy_c}(0) & \varphi_{y_cy_c}(1) & \dots & \varphi_{y_cy_c}(L_{\text{inter}}-1) \\ \varphi_{y_cy_c}(1) & \varphi_{y_cy_c}(0) & \dots & \varphi_{y_cy_c}(L_{\text{inter}}-2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{y_cy_c}(L_{\text{inter}}-1) & \varphi_{y_cy_c}(L_{\text{inter}}-2) & \dots & \varphi_{y_cy_c}(0) \end{pmatrix} \cdot \dots \cdot \begin{pmatrix} h_{\text{inter}}(0) \\ h_{\text{inter}}(1) \\ \vdots \\ h_{\text{inter}}(L_{\text{inter}}-1) \end{pmatrix} = \begin{pmatrix} \varphi_{y_1y_c}(0) \\ \varphi_{y_1y_c}(1) \\ \vdots \\ \varphi_{y_1y_c}(L_{\text{inter}}-1) \end{pmatrix} . \tag{2.24}$$

2.5.2 Joint Optimization

In this case, all the filter coefficients in h_{intra} and h_{inter} are determined jointly. The prediction error is still given by Equation 2.18 and the optimization criterion is defined in Equation 2.17, so that an overall set of equations results that contains all relations between the input signals and the target for the prediction.

Differentiating the expected value with respect to the filter coefficients leads to $L_{\text{intra}} + L_{\text{inter}}$ equations that can be collected in a matrix notation that is (again) similar to the Yule-Walker equations:

$$\mathbf{\Phi} \cdot \mathbf{h} = \boldsymbol{\varphi} \,. \tag{2.25}$$

The matrix and the vectors therein are:

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Phi}_{11} & \mathbf{\Phi}_{1c} \\ \mathbf{\Phi}_{1c}^T & \mathbf{\Phi}_{cc} \end{pmatrix} \tag{2.26}$$

with

$$\mathbf{\Phi}_{11} = \begin{pmatrix} \varphi_{y_1y_1}(0) & \varphi_{y_1y_1}(1) & \dots & \varphi_{y_1y_1}(L_{\text{intra}}-1) \\ \varphi_{y_1y_1}(1) & \varphi_{y_1y_1}(0) & \dots & \varphi_{y_1y_1}(L_{\text{intra}}-2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{y_1y_1}(L_{\text{intra}}-1) & \varphi_{y_1y_1}(L_{\text{intra}}-2) & \dots & \varphi_{y_1y_1}(0) \end{pmatrix} \tag{2.27}$$

$$\mathbf{\Phi}_{1c} = \begin{pmatrix} \varphi_{y_1y_c}(1) & \varphi_{y_1y_c}(0) & \dots & \varphi_{y_1y_c}(2-L_{\text{inter}}) \\ \varphi_{y_1y_c}(2) & \varphi_{y_1y_c}(1) & \dots & \varphi_{y_1y_c}(3-L_{\text{inter}}) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{y_1y_c}(L_{\text{intra}}) & \varphi_{y_1y_c}(L_{\text{intra}}-1) & \dots & \varphi_{y_1y_c}(1+L_{\text{intra}}-L_{\text{inter}}) \end{pmatrix} \tag{2.28}$$

$$\mathbf{\Phi}_{cc} = \begin{pmatrix} \varphi_{y_cy_c}(0) & \varphi_{y_cy_c}(1) & \dots & \varphi_{y_cy_c}(L_{\text{inter}}-1) \\ \varphi_{y_cy_c}(1) & \varphi_{y_cy_c}(0) & \dots & \varphi_{y_cy_c}(L_{\text{inter}}-2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{y_cy_c}(L_{\text{inter}}-1) & \varphi_{y_cy_c}(L_{\text{inter}}-2) & \dots & \varphi_{y_cy_c}(0) \end{pmatrix} \tag{2.29}$$

and

$$\mathbf{h} = \begin{pmatrix} h_{\text{intra}}(1) \\ h_{\text{intra}}(2) \\ \vdots \\ h_{\text{intra}}(L_{\text{intra}}) \\ h_{\text{inter}}(0) \\ h_{\text{inter}}(1) \\ \vdots \\ h_{\text{inter}}(L_{\text{inter}} - 1) \end{pmatrix} \tag{2.30}$$

and

$$\boldsymbol{\varphi} = \begin{pmatrix} \varphi_{y_1 y_1}(1) \\ \varphi_{y_1 y_1}(2) \\ \vdots \\ \varphi_{y_1 y_1}(L_{\text{intra}}) \\ \varphi_{y_1 y_c}(0) \\ \varphi_{y_1 y_c}(-1) \\ \vdots \\ \varphi_{y_1 y_c}(1 - L_{\text{inter}}) \end{pmatrix} \tag{2.31}$$

The inversion of $\boldsymbol{\Phi}$ can be simplified by utilizing the Schur complement $\boldsymbol{\Phi}^C$ of the matrix [Zha10]:

$$\boldsymbol{\Phi}^C = \boldsymbol{\Phi}_{11} - \boldsymbol{\Phi}_{1c} \cdot \boldsymbol{\Phi}_{cc}^{-1} \cdot \boldsymbol{\Phi}_{1c}^T \tag{2.32}$$

To calculate the Schur complement, only the symmetric Toeplitz matrix $\boldsymbol{\Phi}_{cc}$ has to be inverted and the entire matrix can then be inverted by

$$\boldsymbol{\Phi}^{-1} = \begin{pmatrix} \left(\boldsymbol{\Phi}^C\right)^{-1} & -\left(\boldsymbol{\Phi}^C\right)^{-1} \cdot \boldsymbol{\Phi}_{1c} \cdot \boldsymbol{\Phi}_{cc}^{-1} \\ -\boldsymbol{\Phi}_{cc}^{-1} \cdot \boldsymbol{\Phi}_{1c}^T \cdot \left(\boldsymbol{\Phi}^C\right)^{-1} & \boldsymbol{\Phi}_{cc}^{-1} + \boldsymbol{\Phi}_{cc}^{-1} \cdot \boldsymbol{\Phi}_{1c}^T \cdot \left(\boldsymbol{\Phi}^C\right)^{-1} \cdot \boldsymbol{\Phi}_{1c} \cdot \boldsymbol{\Phi}_{cc}^{-1} \end{pmatrix}. \tag{2.33}$$

With this inverse, all filter coefficients can be calculated jointly by

$$\mathbf{h} = \boldsymbol{\Phi}^{-1} \cdot \boldsymbol{\varphi}\,. \tag{2.34}$$

2.5.3 Comparison Between Sequential and Joint Optimization

While all three optimization schemes fundamentally aim at the same target given in Equation 2.17, the different ways of determining the filter coefficients as described in Sections 2.5.1 and 2.5.2 lead to different results as can be seen from the resulting matrices.

The comparison of the three results allows to quantify the gain that is achievable when doing the optimization jointly instead of sequentially or when choosing a better way for the sequential optimization. To do this, the prediction gains of all three setups have to be compared. The prediction gain is defined as the ratio between the energy of the input signal $y_1(k)$ and the energy of the prediction error signal $z_1(k)$.

$$G = \frac{\mathrm{E}\left\{z_1^2(k)\right\}}{\mathrm{E}\left\{y_1^2(k)\right\}} \tag{2.35}$$

Since the input signal $y_1(k)$ of all systems is identical, it is sufficient to compare the energies of the three output signals.

The filter coefficients can not be calculated for infinite filter lengths since the inversion of infinite matrices is not possible in the general case. There are approaches known for the special case of Toeplitz matrices (cf. [BS90]). However, while the individual parts of $\mathbf{\Phi}$ have Toeplitz structure, the Schur complement $\mathbf{\Phi}^C$ has not and since this has to be inverted as well (cf. Equation 2.33), it follows that there is no possibility to determine the filter coefficients for the performance bound, i.e., infinite filter lengths for the predictors.

Hence, some finite special cases will be analyzed in the following:

1. Uncorrelated input signals $y_1(k)$ and $y_c(k)$

   $$\varphi_{y_1 y_c}(\lambda) = 0 \quad \forall \quad \lambda \tag{2.36}$$

2. Scaled signals

   $$y_c(k) = \alpha \cdot y_1(k) \tag{2.37}$$

   This signal property relates to the following correlation properties:

   $$\varphi_{y_1 y_c}(\lambda) = \alpha \cdot \varphi_{y_1 y_1}(\lambda) \quad \forall \quad \lambda \tag{2.38}$$

   $$\varphi_{y_c y_c}(\lambda) = \alpha^2 \cdot \varphi_{y_1 y_1}(\lambda) \quad \forall \quad \lambda \tag{2.39}$$

   with

a) no temporal correlation within the input signals $y_1(k)$ and $y_c(k)$

$$\varphi_{y_1y_1}(\lambda) = 0 \quad \forall \quad \lambda \neq 0 \tag{2.40}$$

b) temporal correlation within the input signals $y_1(k)$ and $y_c(k)$ according to

$$\varphi_{y_1y_1}(\lambda) = \beta^{|\lambda|} \quad \forall \quad \lambda \neq 0 \tag{2.41}$$

This analysis gives a comparison of the energies of the prediction error $\mathrm{E}\left\{z_1^2(k)\right\}$ for the three different optimization strategies. Since these energies lead to lengthy equations even for short filter lengths, the comparison is done here exemplarily for the minimum filter lengths possible: $L_{\text{intra}} = 1$ and $L_{\text{inter}} = 1$. The prediction error $z_1(k)$ in this case can be written as

$$z_1(k) = y_1(k) - h_{\text{intra}}(1) \cdot y_1(k-1) - h_{\text{inter}}(0) \cdot y_c(k) \tag{2.42}$$

for the joint case and the sequential variant that does the intra channel prediction first. When the inter channel prediction is done first for the sequential case, the prediction error can be written as

$$z_1(k) = y_1(k) - h_{\text{intra}}(1) \cdot y_1(k-1) - h_{\text{inter}}(0) \cdot y_c(k) + h_{\text{intra}}(1) \cdot h_{\text{inter}}(0) \cdot y_c(k-1)\,. \tag{2.43}$$

Sequential Optimization – Intra Channel Prediction First

The resulting filter coefficients for the sequential case when doing the intra channel prediction first (cf. the first part of Section 2.5.1) are

$$h_{\text{intra}}(1) = \frac{\varphi_{y_1y_1}(1)}{\varphi_{y_1y_1}(0)} \tag{2.44}$$

and

$$h_{\text{inter}}(0) = \frac{\varphi_{y_1y_1}(0) \cdot \varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1) \cdot \varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0) \cdot \varphi_{y_cy_c}(0)}\,. \tag{2.45}$$

This leads to the following expression for the energy of the prediction error signal (the detailed derivation can be found in Appendix A.1):

$$\begin{aligned}\mathrm{E}\left\{z_1^2(k)\right\} = &\frac{1}{\varphi_{y_1y_1}^2(0) \cdot \varphi_{y_cy_c}(0)} \cdot \big(2\varphi_{y_1y_1}(0) \cdot \varphi_{y_1y_1}(1) \cdot \varphi_{y_1y_c}(0) \cdot \varphi_{y_1y_c}(1) \\ &+ \varphi_{y_1y_1}^3(0) \cdot \varphi_{y_cy_c}(0) - \varphi_{y_1y_1}^2(0) \cdot \varphi_{y_1y_c}^2(0) - \varphi_{y_1y_1}^2(1) \cdot \\ &\cdot \varphi_{y_1y_c}^2(1) - \varphi_{y_1y_1}(0) \cdot \varphi_{y_1y_1}^2(1) \cdot \varphi_{y_cy_c}(0)\big)\,.\end{aligned} \tag{2.46}$$

Sequential Optimization – Inter Channel Prediction First

Doing the inter channel prediction first (cf. the second part of Section 2.5.1) leads to

$$\begin{aligned} h_{\text{intra}}(1) = & \frac{1}{\varphi_{y_1y_1}(0)\cdot\varphi^2_{y_cy_c}(0) - \varphi^2_{y_1y_c}(0)\cdot\varphi_{y_cy_c}(0)}\cdot \\ & \cdot\Big(\varphi_{y_1y_1}(1)\cdot\varphi^2_{y_cy_c}(0) + \varphi^2_{y_1y_c}(0)\varphi_{y_cy_c}(1) \\ & -\varphi_{y_1y_c}(0)\varphi_{y_cy_c}(0)\left(\varphi_{y_1y_c}(1) + \varphi_{y_1y_c}(-1)\right)\Big) \end{aligned} \tag{2.47}$$

and

$$h_{\text{inter}}(0) = \frac{\varphi_{y_1y_c}(0)}{\varphi_{y_cy_c}(0)} \tag{2.48}$$

The energy of the prediction error signal without inserting Equations 2.47 and 2.48 is then (the derivation for this is in Appendix A.2)

$$\begin{aligned} \mathrm{E}\left\{z_1^2(k)\right\} = & \Big(1 + h_{\text{intra}}(1)^2\Big)\cdot\big(\varphi_{y_1y_1}(0) - 2h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(0) \\ & + h_{\text{inter}}(0)^2\cdot\varphi_{y_cy_c}(0)\big) + 2h_{\text{intra}}(1)\cdot\big(h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(1) \\ & + h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(-1) - \varphi_{y_1y_1}(1) - h_{\text{inter}}(0)^2\cdot\varphi_{y_cy_c}(1)\big) \end{aligned} \tag{2.49}$$

Joint Optimization

Using the joint optimization from Section 2.5.2 leads to

$$h_{\text{intra}}(1) = \frac{\varphi_{y_1y_1}(1)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi^2_{y_1y_c}(1)} \tag{2.50}$$

and

$$h_{\text{inter}}(0) = \frac{\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi^2_{y_1y_c}(1)}\,. \tag{2.51}$$

In this case, the energy of the prediction error is (the detailed derivation is given in Appendix A.3)

$$\begin{aligned}
\mathrm{E}\left\{z_1^2(k)\right\} = & \frac{1}{\left(\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)\right)^2}\cdot\Big(\varphi_{y_1y_1}^3(0)\cdot\varphi_{y_cy_c}^2(0) \\
& - \varphi_{y_1y_1}^2(0)\cdot\varphi_{y_1y_c}^2(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_1}^2(1)\cdot\varphi_{y_cy_c}^2(0) \\
& + \varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}^4(1) - 2\varphi_{y_1y_1}^2(0)\cdot\varphi_{y_1y_c}^2(1)\cdot\varphi_{y_cy_c}(0) \\
& + \varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}^2(0)\cdot\varphi_{y_1y_c}^2(1) + 3\varphi_{y_1y_1}^2(1)\cdot\varphi_{y_1y_c}^2(1)\cdot\varphi_{y_cy_c}(0) \\
& - 2\varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}^3(1) - 2\varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}^3(1)\cdot\varphi_{y_cy_c}(0) \\
& + 2\,\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}^2(1)\cdot\varphi_{y_cy_c}(0)\Big)
\end{aligned} \tag{2.52}$$

Comparison between the Optimization Strategies

The quotients of the three energies from Equations 2.46, 2.49 and 2.52 can be utilized to gain some insight into the behaviour of the different optimization strategies for the aforementioned special cases. Three quotients can be defined as

$$G_1 = \frac{\mathrm{E}\left\{z_1^2(k)\right\}_{\text{sequential,intra first}}}{\mathrm{E}\left\{z_1^2(k)\right\}_{\text{sequential,inter first}}} \tag{2.53}$$

$$G_2 = \frac{\mathrm{E}\left\{z_1^2(k)\right\}_{\text{sequential,intra first}}}{\mathrm{E}\left\{z_1^2(k)\right\}_{\text{joint}}} \tag{2.54}$$

$$G_3 = \frac{\mathrm{E}\left\{z_1^2(k)\right\}_{\text{sequential,inter first}}}{\mathrm{E}\left\{z_1^2(k)\right\}_{\text{joint}}} \tag{2.55}$$

There are no simplifications possible in any of the three quotients of two energies, and since the full equations are very lengthy while simultaneously not offering much insight, the complete formulas are omitted here.

All quotients and the filter coefficients are calculated for the different finite cases previously defined:

1. **No correlation between the input signals $y_1(k)$ and $y_c(k)$**
   Inserting Equation 2.36 into Equations 2.53, 2.54 and 2.55 leads to

$$G_1 = G_2 = G_3 = 1\,. \tag{2.56}$$

As could be expected, there is no gain from the inter channel prediction in this case and hence the gain for all optimization schemes is identical. The same finding would also be possible when looking at the filter coefficients $h_{\text{intra}}(1)$ (cf. Equations 2.44 and 2.50) and $h_{\text{inter}}(0)$ (cf. Equations 2.45 and 2.51). All three schemes lead to identical filter coefficients, namely

$$h_{\text{intra}}(1) = \frac{\varphi_{y_1 y_1}(1)}{\varphi_{y_1 y_1}(0)} \tag{2.57}$$

and

$$h_{\text{inter}}(0) = 0\,. \tag{2.58}$$

2. **Fully correlated (only scaled) signals**
Combining the gain quotients with Equations 2.38 and 2.39 and separately analyzing the two variants thereof allows to gain further insights into the behaviour of the system.

   a) With Equation 2.40, again

   $$G_1 = G_2 = G_3 = 1 \tag{2.59}$$

   results and, of course, the filter coefficients are also identical:

   $$h_{\text{intra}}(1) = 0 \tag{2.60}$$

   and

   $$h_{\text{inter}}(0) = \frac{1}{\alpha}\,. \tag{2.61}$$

   b) The variant with fixed temporal correlation within the input signals $y_1(k)$ and $y_c(k)$ as described in Equation 2.41 leads to different results. Comparing the sequential variant that does the inter channel prediction first and the joint optimization scheme, respectively, with the sequential variant which performs the intra channel prediction first gives

   $$G_1 = G_2 = \infty\,. \tag{2.62}$$

   The filter coefficients in this case are different, for the sequential case with intra channel prediction first, they are

   $$h_{\text{intra}}(1) = 0 \tag{2.63}$$

and

$$h_{\text{inter}}(0) = \frac{1}{\alpha} \cdot \frac{\varphi^2_{y_1 y_1}(0) - \beta^2}{\varphi^2_{y_1 y_1}(0)}\,, \tag{2.64}$$

for the sequential case with inter channel prediction first and the joint case, they are

$$h_{\text{intra}}(1) = 0 \tag{2.65}$$

and

$$h_{\text{inter}}(0) = \frac{1}{\alpha}\,. \tag{2.66}$$

Since their filter coefficients are identical, the final two case also perform identically and

$$G_3 = 1 \tag{2.67}$$

results.

Taking all setups into account, it has to be stated that both sequential optimization setups have the advantage when it comes to computational complexity that the determination of the filter coefficients is based on smaller matrices which additionally have Toeplitz structure making it even easier to invert them. However, the performance of the resulting system in the sequential cases depends on the exact system setup.

In the exemplary setup that was analyzed here, the sequential setup which does the inter channel prediction first performs as good as the joint optimization setup (cf. Eqs. 2.56, 2.59, 2.67). For arbitrary signals and filter lengths however, only the joint optimization scheme can guarantee the optimum performance and it is up to specific parameters of the system design if the possible additional gain of the joint scheme outweighs its increased complexity.

2.6 A Flexible Structure for Multi Channel Linear Prediction

In the previous sections, different variants of integrating intra and inter channel prediction were devised and analyzed. One part was fixed in all those variants:

Only the most recent samples from the different input channels were utilized. In this section, a more flexible system is devised which distributes the available filter taps between intra and inter channel prediction while simultaneously having an increased search range, i.e., not only the most recent samples but those samples which lead to the largest prediction gain are chosen.

Usually, the number of coefficients for intra channel prediction is fixed and set to (depending on the sampling rate) values of about 8 to 16 (cf. the standardized speech codecs in [3GP88, ITU96c, ETS09]). The argument for this number of coefficients is usually the length of the human vocal tract which can be modeled as an IIR filter [VM06] and a compromise between the complexity and the achievable prediction gain. The number of coefficients for the inter channel prediction can not be motivated by a similar model decision in general. If a coding system for a certain microphone setup shall be parametrized, the maximum length of the impulse response between the microphones could be used as a reference for the length of the inter channel predictor. Since the length of the impulse response is highly dependent on the acoustic environment of the microphone setup, this is only possible for very specific hardware setups, e.g., a hands-free unit that is built into a car. As a simple approach, it might be possible for known distances between the microphones, to parametrize the inter channel predictor to account at least for the direct path of the impulse response between the microphones.

Alternatively, using a signal alignment scheme as the first step in the signal processing system is possible such that the delay between the input signals to the predictive coding (i.e., $y_1(k)$ and $y_c(k)$) is compensated. This way, the most strongly correlated parts of the signals are aligned and the maximum prediction gain can be ensured. Additionally, the most dominant parts in the impulse response between the two microphones are (in many realistic environments, cf. Section 2.2) in the close vicinity of the strongest component.

The classical approach to linear prediction would be to define some optimization criterion (cf. Section 2.4.2) for the prediction error and to calculate filter coefficients accordingly.

A novel, alternative way for the determination of prediction filter coefficients is presented here. Therein, the prediction procedure is interpreted in a frame-wise manner as a stepwise model building procedure which offers additional degrees of freedom and is related to LTP.

In the following, the overall system is introduced first and different ways to determine these filter taps are presented and analyzed. The basic principle will be shown to be closely related to a known concept from the area of generic signal decomposition.

2.6.1 Signals and Variables

The basic system that is used for the reformulation of the multichannel predictive system is depicted in Figure 2.7 for two out of M channels. The extension to the multi channel case is straightforward: All instances of $y_c(k)$ are replaced by multiple signals, everything else remains identical to the presented two channel case.

A joint filter structure is used which adaptively picks the most useful past signal samples of both $y_1(k)$ and/or $y_c(k)$ by means of a coefficient vector $\mathbf{b}$.

The signal processing in this system is done in a blockwise manner with non-overlapping blocks. Every block has a length of T_{fl} and hence consists of $N = f_s \cdot T_{fl}$ samples.

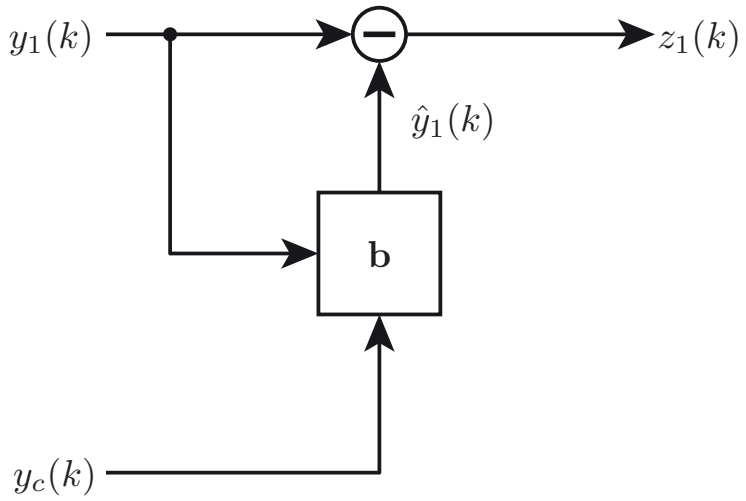


Figure 2.7: Flexible intra and inter channel prediction system

The estimate $\hat{y}_1(k)$ of the input signal $y_1(k)$ is based both on $y_1(k)$ itself and $y_c(k)$. To predict the current sample of $y_1(k)$, all but the current sample of $y_1(k)$ (i.e., $y_1(k-\lambda)\ \forall\ \lambda \geq 1$) and all samples of $y_c(k)$ (i.e., $y_c(k-\lambda)\ \forall\ \lambda \geq 0$) are available.

In any practical environment, a number of samples $N_{lb,i}$ for the maximum look-back for both signals $y_1(k)$ and $y_c(k)$ has to be predefined when designing the prediction system. This is inherently done within regular LP systems as well by setting the length L of the prediction filter. With this value set, a matrix of excerpts is constructed from the past of $y_1(k)$ and the past and the present of $y_c(k)$ with dimensions $N \times (N_{lb,1} + N_{lb,c} + 1)$.

For the following derivations, a vectorized notation is introduced which is based on vectors of N consecutive values of the signal $y_1(k)$ or $y_c(k)$:

$$\mathbf{y}_{c,k} = \Big(y_c(k) \ \ldots\ y_c(k-N+1) \Big)^T \ . \tag{2.68}$$

At the very beginning, these vectors are padded with zeros as long as $k < N_{lb,c} + N$.

Both the past of the signal to be predicted and the other input signal can then be collected in matrix form as

$$\mathbf{\Lambda}_{\text{intra}} = \left(\mathbf{y}_{1,k-N_{lb,1}} \cdots \mathbf{y}_{1,k-1} \right) \tag{2.69}$$

and

$$\mathbf{\Lambda}_{\text{inter}} = \left(\mathbf{y}_{c,k-N_{lb,c}} \cdots \mathbf{y}_{c,k} \right) , \tag{2.70}$$

respectively.

The signal that shall be predicted (i.e., the current frame of $y_1(k)$) can be written in this notation as

$$\mathbf{y}_{1,k} = \left(y_1(k) \ldots y_1(k-N+1) \right)^T . \tag{2.71}$$

Since both signals can be used within the prediction process, this complete basis for the prediction can be arranged in a joint matrix as

$$\mathbf{\Lambda} = [\mathbf{\Lambda}_{\text{intra}} \, \mathbf{\Lambda}_{\text{inter}}] . \tag{2.72}$$

In the case of the classical single channel prediction, $\mathbf{\Lambda}$ only contains the $N_{lb,1}$ vectors from the past of $y_1(k)$.

The target of the prediction scheme is to find the L coefficients that lead to an optimum filter with respect to the optimization criterion. This can be expressed as a multiplication of the matrix $\mathbf{\Lambda}$ (see Equation 2.72) with a vector $\mathbf{b}$ to get an estimate $\hat{\mathbf{y}}$ of the prediction target $\mathbf{y}$ (see Eq. 2.71).

$$\hat{\mathbf{y}}_{1,k} = \mathbf{\Lambda} \cdot \mathbf{b} \tag{2.73}$$

The estimate contains the N recent values of $\hat{y}_1(k)$ according to

$$\hat{\mathbf{y}}_{1,k} = \left(\hat{y}_1(k) \ldots \hat{y}_1(k-N+1) \right)^T . \tag{2.74}$$

The prediction coefficient vector $\mathbf{b}$ is a column vector of length $(N_{lb,1} + N_{lb,c} + 1)$ that contains no more than L non-zero entries. The determination of the filter coefficients is based on the auto correlation $\varphi_{y_1 y_1}(\lambda)$

and the cross correlation $\varphi_{y_1 y_c}(\lambda)$. The necessary values can be calculated by means of a matrix multiplication. A vector $\boldsymbol{\varphi}_{y_1 y_1, y_1 y_c}$ of correlation values results according to:

$$\boldsymbol{\varphi}_{y_1 y_1, y_1 y_c} = \mathbf{y}_{1,k}^T \cdot \boldsymbol{\Lambda} \tag{2.75}$$

This row vector contains $2 \cdot N_{lb} + 1$ auto correlation and cross correlation values:

$$\boldsymbol{\varphi}_{y_1 y_1, y_1 y_c} = \left(\varphi_{y_1 y_1, y_1 y_c}(0)\ \varphi_{y_1 y_1, y_1 y_c}(1)\ \ldots\ \varphi_{y_1 y_1, y_1 y_c}(2 \cdot N_{lb}) \right) \tag{2.76}$$

The mapping of values of $\varphi_{y_1 y_1}(\lambda)$ and $\varphi_{y_1 y_c}(\lambda)$ onto $\varphi_{y_1 y_1, y_1 y_c}(\lambda)$ is done according to

$$\varphi_{y_1 y_1, y_1 y_c}(\lambda) = \begin{cases} \varphi_{y_1 y_1}(\lambda - N_{lb}) & 0 \leq \lambda < N_{lb} \\ \varphi_{y_1 y_c}(\lambda - 2 \cdot N_{lb}) & N_{lb} \leq \lambda \leq 2 \cdot N_{lb} \end{cases} \tag{2.77}$$

With this basic setup, the filter coefficients b_i and the delays D_i can be determined in different ways, two novel approaches are presented here:

- A stepwise statistical modeling procedure based on [TF06] in Section 2.6.2
- A stepwise correlation-based concept in Section 2.6.3 to determine
  - the delays *and* the filter coefficients sequentially (i.e., first, b_1 and D_1, then b_2 and D_2, and so on)
  - all delays sequentially and then all filter coefficients jointly.

Irrespective of the way how the remaining coefficients and the delays are determined, the first values can be calculated directly from $\boldsymbol{\varphi}_{y_1 y_1, y_1 y_c}(\lambda)$ by searching the maximum of the correlation and determining the coefficient by dividing the energy of the signal segment that shall be predicted by the correlation at the point D_1:

$$D_1 = \arg\max_{\lambda} \varphi_{y_1 y_1, y_1 y_c}(\lambda) \tag{2.78}$$

$$b_1 = \frac{\mathbf{y}_{1,k}^T \cdot \mathbf{y}_{1,k}}{\varphi_{y_1 y_1, y_1 y_c}(D_1)} \tag{2.79}$$

After this first step, the prediction coefficient vector has only one entry b_1 at position D_1.

$$\mathbf{b}_1 = \left(\underbrace{0 \ldots 0}_{D_1 - 1} \quad b_1 \quad \underbrace{0 \ldots 0}_{N_{lb,1} + N_{lb,c} + 1 - D_1} \right)^T \tag{2.80}$$

The remaining $L-1$ coefficients can be determined by different methods which are described in the following.

2.6.2 Stepwise Statistical Regression

The statistical linear prediction is based on the *stepwise statistical regression* which is a well-known tool for modeling unknown relations between input and output values of a unknown system [TF06]. It basically consists of 4 steps:

1. Initialize the input and output variables.
2. Try to add another non-zero entry in $\mathbf{b}$ based on a statistical test.
3. Try to remove an unnecessary entry in $\mathbf{b}$, again based on a statistical test.
4. Check if enough non-zero entries in $\mathbf{b}$ are present (i.e., if $|\mathbf{b}| = L$ with $|\cdot|$ denoting the cardinality of $\cdot$).
   - If yes, the necessary coefficients were found and the procedure can be aborted.
   - If no, return to step 2 and continue.

In certain cases when the prediction is already perfect with less than L coefficients, i.e., $\mathbf{y}_{1,k} - \hat{\mathbf{y}}_{1,k} = 0 \; \exists \; |\mathbf{b}| < L$, this algorithm does not find L coefficients and hence would not finish. A simple additional check has to be carried out in step four of the algorithm to identify these cases.

With the aforementioned starting point, the statistical regression is efficiently initialized. The results with this prediction paradigm will be presented in Section 2.6.7 in comparison to the alternative approach that is described in the following section.

2.6.3 Stepwise Correlation-based Regression

The second approach can also be derived based on the matrix formulation in Eq. 2.73. The decisive change is the way to determine the vector $\mathbf{b}$: A correlation-based approach is used here instead of the statistical tests that were utilized before.

No matter if the filter coefficients b_i are determined in a stepwise manner or jointly, the delays D_i have to be determined first. The first delay D_1 is determined in analogy to Eq. 2.78 as the maximum of the matrix product between the signal to be predicted $\mathbf{y}_{1,k}$ and the basis for the prediction $\mathbf{y}$. To determine the second delay, the residual signal $\mathbf{y}'_{1,k}$ after the first step has to be predicted. Hence, also the first prediction coefficient b_1 has to be determined in analogy to Eq. 2.79 to calculate $\mathbf{y}'_{1,k}$ (with the prediction coefficient vector $\mathbf{b}_1$ as defined in Eq. 2.80):

$$\mathbf{y}'_{1,k} = \mathbf{y}_{1,k} - \mathbf{\Lambda} \cdot \mathbf{b}_1 \tag{2.81}$$

With this new target for the prediction, the next delay can be determined in a similar manner by searching for the maximum correlation between the basis for the prediction and the prediction error.

$$D_2 = \arg\max_{\lambda} \boldsymbol{\varphi}_{yy'}(\lambda) \tag{2.82}$$

In order to calculate the new prediction target, the residual signal $\mathbf{y}''$ of the second step, a second prediction coefficient b_2 has to be determined by

$$b_2 = \frac{\varphi_{yy}(0)}{\varphi_{yy}(D_2)} \tag{2.83}$$

so that a new prediction coefficient vector can be written as

$$\mathbf{b}_2 = \left(\underbrace{0 \ldots 0}_{D_1-1} \quad b_1 \quad \underbrace{0 \ldots 0}_{N_{lb}-D_1} \right)^T + \left(\underbrace{0 \ldots 0}_{D_2-1} \quad b_2 \quad \underbrace{0 \ldots 0}_{N_{lb}-D_2} \right)^T . \tag{2.84}$$

This procedure is repeated until L delays D_i have been found.

For the stepwise determination of the filter coefficients, only one more step is necessary since all filter coefficients but one were already calculated in the process of finding the delays and only the last filter coefficient b_L is missing. This can be done in analogy to Eq. 2.83 and the final prediction error signal can be determined by

$$\mathbf{z} = \mathbf{y}_{1,k} - \mathbf{\Lambda} \cdot \mathbf{b} \tag{2.85}$$

To jointly determine all filter coefficients at once, a process very similar to the determination of the filter coefficients in classical single channel prediction can be used. The target is to minimize the energy of the prediction error signal (cf. Equation 2.85).

$$\left(\mathbf{z}^T \cdot \mathbf{z}\right) \rightarrow \min \tag{2.86}$$

This can be done by differentiating $\mathbf{z}^T \cdot \mathbf{z}$ with respect to the filter coefficients b_λ and setting the result equal to zero.

$$\frac{\mathrm{d}\mathbf{z}^T \cdot \mathbf{z}}{\mathrm{d}\mathbf{b}} = -2\mathbf{\Lambda}^T \cdot \mathbf{y}_{1,k} + 2\mathbf{\Lambda}^T \left(\mathbf{\Lambda} \cdot \mathbf{b}\right) \stackrel{!}{=} 0 \tag{2.87}$$

With the Moore-Penrose inverse $\cdot^\dagger$ [Moo20, PT55], this can be solved and gives

$$\mathbf{b} = \mathbf{\Lambda}^\dagger \cdot \mathbf{z}\,. \tag{2.88}$$

Differences in performance between the stepwise and the joint determination of the filter coefficients will be analyzed in Section 2.6.7.

Another point is worth mentioning with both stepwise procedures: They have been presented here in a setup that aims for a certain number of coefficients. Additionally, a target prediction gain can be defined as well and as many indices and coefficients are identified as are necessary to reach this prediction gain.

2.6.4 Relation to Matching Pursuit

Matching Pursuit is a procedure that was proposed in [MZ93] to decompose signals into linear summations of so-called atoms. There are different dictionaries for these atoms that have been proposed and used in different scenarios. In speech coding in particular, sinusoidal dictionaries have been used in, e.g., [ECG00] where also a dynamic dictionary adaptation was proposed.

The dictionary in the presented approaches is given by the matrix $\mathbf{\Lambda}$ which is highly adaptive to the current status of the two signals. The presented approach can hence be understood both as a novel highly adaptive variant of Matching Pursuit as well as a novel take on predictive coding systems.

2.6.5 Packet Losses

So far, only the transmitting side of the communication system has been considered. In a realistic application of the presented concepts, the signals also have to be transmitted to the receiving side. This transmission introduces additional challenges due to errors that can occur, e.g., on a radio link or due to a loss of data, e.g., in *Internet Protocol* (IP) networks.

With the dynamic construction of the dictionary from the past of the signal, packet losses are very critical. A single lost packet could lead to error propagation to all subsequent packets.

In IP networks, packet losses can be fairly frequent depending on the circumstances. Especially the delay constraints of the application are an important aspect in this regard since packets are often not really lost but arrive too late due to jitter in the transmission link. Some systems even include mechanisms to ensure that faulty or missing packets are sent again, e.g, the *Long Term Evolution* (LTE) physical layer features a so-called *Hybrid Automatic Repeat reQuest* (HARQ) system which forces retransmissions of not correctly received packets (HARQ Type I) or the transmission of additional parity bits (HARQ TYpe II). This can decrease the bit error rate at the expense of increasing the necessary delay.

In the unified approach, there is a decisive difference between packets arriving too late and packets not arriving at all. Really lost packets can lead to infinite error propagation while late packets lead to a significantly less dramatic effect. Common error concealment techniques can be used in that case to mitigate the detrimental effect of the missing packet while a later resynchronization of the decoder is still possible once the packet has arrived.

2.6.6 Interpretation as Generalized Long Term Prediction

In single channel linear prediction, *Long Term Prediction* (LTP) (cf. Section 2.4.1) refers to a second prediction step which is carried out after the short term correlation within the signal was removed by the first prediction step as described in Section 2.4. Regular LTP can be described as a special case of the proposed approach by reducing the number of determined filter taps L to one (or three in the case of an interpolating LTP [ITU96b]) and only using $y_1(k)$ as the basis for the prediction.

The new approach offers additional degrees of freedom that can be used to increase the achievable prediction gain and thus decrease the data rate necessary for the transmission. The resulting structure with the new approach is identical to having multiple independent LTP filters in parallel.

2.6.7 Simulation Example

The vectors of filter coefficients $\mathbf{h}_{\text{intra}}$ and $\mathbf{h}_{\text{inter}}$ are determined both by the stepwise statistical regression from Section 2.6.2 and the stepwise correlation approach from Section 2.6.3 to compare the performance of the different approaches.

The respective approaches are used in the way that was already described earlier and a maximum number of three coefficients are determined. As a comparison, a regular intra channel linear prediction with 10 coefficients is utilized.

The different setups are simulated for the 3GPP audio dataset [3GP07] consisting of approximately ten minutes of stereo signals at a sampling frequency of $f_s = 48\,\text{kHz}$. The signals in the dataset contain clear and noisy speech from various talkers in different languages as well as music signals. By either using the left channel $y_l(k)$ as the signal to be predicted $y_1(k)$ and the right channel $y_r(k)$ as the additional basis for the prediction $y_c(k)$ or vice versa, the effective length of the dataset for this evaluation is easily doubled. The processing is done on frames of 20 ms.

The results are quantified based on the achieved prediction gains:

$$G_p(l) = \frac{\mathrm{E}\left\{y_l^2(k)\right\}}{\mathrm{E}\left\{z_l^2(k)\right\}} \qquad G_p(r) = \frac{\mathrm{E}\left\{y_r^2(k)\right\}}{\mathrm{E}\left\{z_r^2(k)\right\}} \tag{2.89}$$

The aforementioned two variants of using the two channels of the signals in the dataset are here denoted by $G_p(l)$ and $G_p(r)$. The prediction gain is calculated for both channels of the original signal as the signal to be predicted and then averaged

$$G_p = \frac{G_p(l) + G_p(r)}{2} \tag{2.90}$$

For the baseline setup of only intra channel linear prediction (with a filter length of $L_{\text{intra}} = 10$) as it is found in many known speech codecs, an average prediction gain of $10 \cdot \log_{10}(G_p) = 19.78\,\text{dB}$ results.

The stepwise statistical model from Section 2.6.2 leads to an average prediction gain of $10 \cdot \log_{10}(G_p) = 36.58\,\text{dB}$. This is practically identical to the performance of the stepwise correlation approach from Section 2.6.3 which achieves an average prediction gain of $10 \cdot \log_{10}(G_p) = 36.24\,\text{dB}$ for the sequential determination of the filter coefficients and $10 \cdot \log_{10}(G_p) = 36.39\,\text{dB}$ for the joint determination of the filter coefficients, respectively.

All three variants perform similar but a large difference in computational complexity especially between the statistical modeling and the correlation-based concept can be observed. The difference between the stepwise and the joint determination of the filter coefficients is almost negligible so that the additional step in Equation 2.88 for the joint determination is not necessary and can be omitted.

2.7 Conclusions

This chapter has introduced the notation and the fundamental background for the signal processing systems that will be devised in the following chapters. This notation was presented along with some analyses of the acoustic environment that is shown to be an important aspect of any speech and audio signal processing or signal transmission system that shall be used in a realistic environment.

In the area of linear predictive coding techniques, a short overview on single channel linear predictive coding was followed by a novel way for the calculation of filter taps. The new procedure is a generalization of Long Term Prediction and can also be interpreted as a highly adaptive way of Matching Pursuit. This novel setup allows to use linear prediction in a much more flexible manner. Especially in the multi channel case, this reformulation was shown to lead to a significantly improved prediction gain.

3

Outer Stage – Preconditioning and Enhancement

The two parts of the outer stage of the multichannel signal processing system is depicted in Figure 3.1 (cf. the system overview in Section 1.2) and described in this chapter. It is a network of *Finite Impulse Response* (FIR) filters that can be used for different tasks. Two different possibilities each for the transmitting and the receiving side are discussed here.

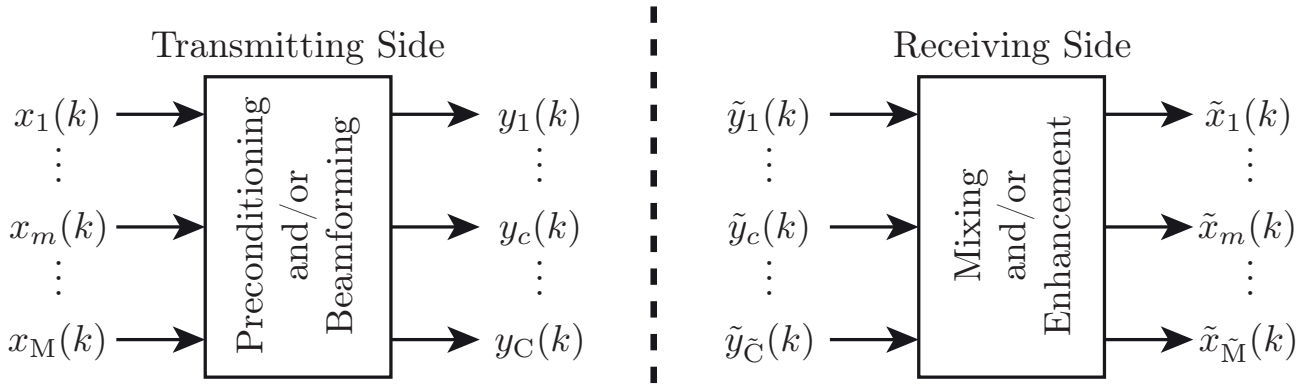


Figure 3.1: Outer stage of the multi channel signal processing system

On the transmitting side, the two applications are

- Preconditioning the signals by means of appropriate channel mixing schemes for the predictive coding which itself will be covered in Chapter 4,
- Beamforming to extract audio information from a certain spatial position.

On the receiving side, a channel mixing can be used to invert the preconditioning from the transmitting side and additionally, a scheme for improving speech intelligibility for coded transmissions is presented as well.

In general and independent from the task, the channel mixing approaches for the outer stage can be subdivided into two classes:

- Fixed filter coefficients that are determined during system design,
- Filter coefficients that are adapted to the current state of the system and the current input signals.

Both classes are advantageous for certain applications. Fixed designs as presented in Section 3.2.1 are very useful as they can exhibit synergies when combined with the inner stage which will be covered in Chapter 4. The adaptive setups in Section 3.2.2 are important in cases where no synergies with the inner stage shall be exploited.

The transmitting side of a generic mixing system is depicted in Figure 3.2. Its inputs are the signals $x_1(k) \dots x_{\mathrm{M}}(k)$ and the outputs are the signals $y_1(k) \dots y_{\mathrm{C}}(k)$. This mixing system exhibits strong similarities to the convolutive mixture model that was introduced in Section 2.1 in that its output signals are summations of the filtered input signals. Between each microphone and each summation point, there is an FIR filter $f_{m,c}(k)$ of arbitrary length connecting microphone m and summation point c. Hence, the signal $y_c(k)$ at summation point c can be calculated as the sum of all filtered microphone signals:

$$y_c(k) = \sum_{m=1}^{\mathrm{M}} f_{m,c}(k) * x_m(k) \tag{3.1}$$

The corresponding system at the receiving side looks very similar only that it has C input signals $\tilde{y}_1(k) \dots \tilde{y}_{\mathrm{C}}(k)$ and M output signals $\tilde{x}_1(k) \dots \tilde{x}_{\mathrm{M}}(k)$. The filters there are also FIR filters $\tilde{f}_{c,m}(k)$ so that the calculation of the output signals is done in analogy to Equation 3.1 by:

$$\tilde{x}_m(k) = \sum_{c=1}^{\tilde{\mathrm{C}}} \tilde{f}_{c,m}(k) * \tilde{y}_c(k) \tag{3.2}$$

One fundamental target of any mixing scheme on the transmitting as well as on the receiving side is to match the number of channels to the number of inputs of the next element in the signal processing chain.

If this stage shall be used as a preprocessing for a subsequent transmission system, two possibilities can arise. The target may be to decrease the number of

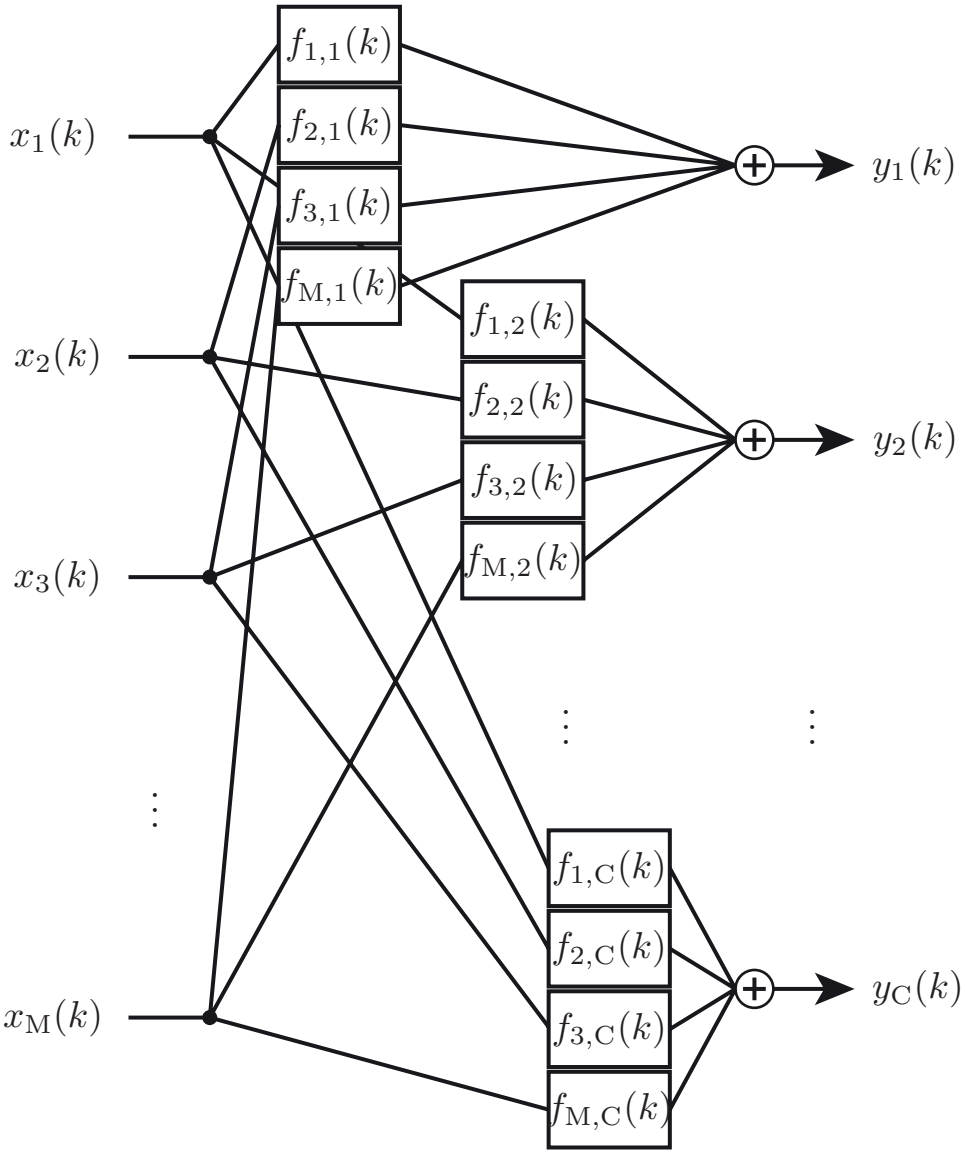


Figure 3.2: Generic model for the outer stage at the transmitting side

input channels if it is too large for the capabilities of the transmission system (i.e., $\mathrm{C} < \mathrm{M}$), then, this stage is commonly called downmixing. The downmixing is one core element of many multi-channel coding schemes. The other possibility may be that the transmission system relies on certain properties of its input signals (e.g., the transmission system that is presented in Chapter 4 needs one channel to be a normalized sum of all input channels). Then, the mixing stage has to ensure that these properties are present in its output.

An alternative view on the outer stage of the multi-channel transmission system can be gained from looking at beamforming, where the principle of using numerous microphone signals to generate one or several output channels can also be found. A novel concept for the determination of the weighting coefficients for weighted delay-and-sum microphone arrays was introduced in [SHWV12]

and further refined for filter-and-sum microphone arrays in [HSWV13]. The concept is based on a numerical optimization of the reception characteristic of the microphone array. The optimization procedure is shown to improve the reception characteristic in such a way that it closely approximates a target which can be defined according to the application.

The two basic setups mixing and beamforming can also be distinguished by the receiver side processing. While beamforming is carried out completely on the transmitting side, there is a suitable reconstruction stage necessary for (down)mixing schemes where an additional constraint could be to reconstruct the original (microphone) signals $x_1(k) \dots x_{\mathrm{M}}(k)$ by decoding. On the receiving side, a scheme that aims at increasing the speech intelligibility can be integrated into this structure as well.

3.1 Including the Acoustic Environment

The filtered sum of all input channels is utilized here which requires $\mathrm{M} \cdot \mathrm{C}$ filters $f_{m,c}(k)$ from microphone m to internal channel c. Depending on the specific application, it is useful to also include the acoustic environment in the system analysis and design. Starting at a certain source location with index n in the acoustic environment of the microphone, the signal $x_m(k)$ at microphone m is (cf. Eq. 2.5)

$$x_m(k) = h_{n,m}(k) * s_n(k). \tag{3.3}$$

Inserting these microphone signals into Eq. 3.1, the signal at the summation point can also be written as:

$$y_c(k) = \sum_{m=1}^{\mathrm{M}} f_{m,c}(k) * \left(h_{n,m}(k) * s_n(k)\right). \tag{3.4}$$

With these filters and using the associative property of the convolution, an overall filter $g_{n,c}(k)$ from each source point n to every internal channel c can be determined as the cascade of the *Room Impulse Responses* (RIRs) $h_{n,m}(k)$ and the path filters $f_{m,c}(k)$

$$g_{n,c}(k) = \sum_{m=1}^{\mathrm{M}} h_{n,m}(k) * f_{m,c}(k) \tag{3.5}$$

The specific determination of the filters $f_{1,1}(k) \dots f_{\mathrm{M,C}}(k)$ depends on the exact task that should be fulfilled with these filters. Different design rules will be presented in the following sections.

3.2 Channel Mixing

3.2.1 Fixed Approaches

Depending on the recording and transmission setup, certain fixed filter setups are advantegeous. These fixed setups are known throughout the transmission system. Hence, every signal processing step can rely on this knowledge and no additional information about the utilized downmixing process has to be provided. This allows these approaches to provide some coding gain without transmitting any additional data at all.

Normalized Summation

One of the most obvious concepts for downmixing two channels to one channel is the normalized summation of the two input channels $x_1(k)$ and $x_2(k)$ to generate one transmission signal $y_1(k)$:

$$y_1(k) = \frac{x_1(k) + x_2(k)}{2}\,. \tag{3.6}$$

This concept is commonly used as part of the so-called sum-difference encoding which will be treated in the next section. This downmixing scheme can also be generalized for more than two input channels by:

$$y_1(k) = \frac{\sum_{m=1}^{\mathrm{M}} x_m(k)}{\mathrm{M}}\,. \tag{3.7}$$

All the mixing filters $f_{m,1}(k)$ in this case are identical:

$$f_{1,1}(k) = f_{2,1}(k) = \cdots = f_{\mathrm{M},1}(k) = \begin{cases} \frac{1}{\mathrm{M}} & k = 0 \\ 0 & k \neq 0\,. \end{cases} \tag{3.8}$$

The multi channel case is not considered explicitly in the following steps since all findings can already be made for the two channel case (i.e., $\mathrm{M} = 2$) as in Eq. 3.6.

This system has one major weakness: it is very sensitive to phase relations between the two input signals. If a phase difference between the two signals of π is present (i.e., $x_2(k) = -x_1(k)$) it follows that

$$y_1(k) = \frac{x_1(k) + x_2(k)}{2} = 0. \tag{3.9}$$

Of course, this is an extreme (and unrealistic) special case but even in the general case, this simple summation concept leads to an overall impulse response $g_1(k)$ from a single source emitting the signal $s(k)$ to the single transmission channel which itself is the normalized summation of the RIRs from the source to the two microphones:

$$\begin{aligned} y_1(k) &= \frac{x_1(k) + x_2(k)}{2} \\ &= \frac{h_1(k) * s(k) + h_2(k) * s(k)}{2} \\ &= \frac{h_1(k) + h_2(k)}{2} * s(k) \\ &\Rightarrow g_1(k) = \frac{h_1(k) + h_2(k)}{2} \,. \end{aligned} \tag{3.10}$$

Hence, even for the simple delay and damping model (cf. Equation 2.1), a comb filter results which exhibits a distinct frequency dependency with strong notches and peaks.

Sum and Difference Coding for Multi-Channel Signals

The use of sum and difference coding is a well-known technique in the area of stereo coding. Therein, the two input channels $x_1(k)$ and $x_2(k)$ are combined in a butterfly structure as depicted in Figure 3.3.

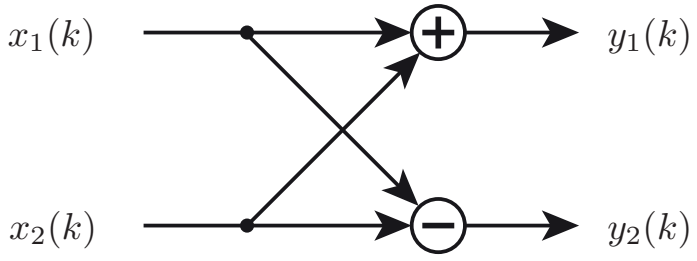


Figure 3.3: Fixed mixing stage of sum and difference coding

This leads to one normalized sum (see Eq. 3.6) $y_1(k)$ and one normalized difference $y_2(k)$.

$$y_2(k) = \frac{x_1(k) - x_2(k)}{2} \,. \tag{3.11}$$

Analogous to Equation 3.8, the mixing filters for the second channel $f_{n,2}(k)$ are identical absolute value with different sign:

$$f_{1,2}(k) = -f_{2,2}(k) = \begin{cases} \frac{1}{2} & k = 0 \\ 0 & k \neq 0 \end{cases} \tag{3.12}$$

Coding gains in this setup are possible if there are no major phase and level differences between the channels as then there will be a concentration of the signal energy in $y_1(k)$ and the energy of $y_2(k)$ will be significantly lower. This energy difference can then be exploited by subsequent coding stages (e.g., by simply utilizing different quantizers for the two signals).

3.2.2 Adaptive Approaches

The adaptive approaches have the edge over the fixed approaches when it comes to the possible prediction gain at the expense of additional data rate to transmit the adaptive parameters that are determined in the sender. Alternatively, there is also the possibility of backward adaptation of the filter coefficients which is a combination of the best of both worlds as long as no transmission errors occur. As soon as these errors occur though, a misalignment between encoder and decoder results which may dramatically influence the overall transmission quality. In the following, no backwards adaptive approaches are considered.

Generalized Sum and Difference

This approach is a similar to the fixed version that is presented in Section 3.2.1: From the two input channels $x_1(k)$ and $x_2(k)$, a sum signal $y_1(k)$ and a difference signal $y_2(k)$ are calculated according to

$$y_1(k) = \frac{\alpha \cdot x_1(k) + (2-\alpha) \cdot x_2(k)}{2}\,. \tag{3.13}$$

and

$$y_2(k) = \frac{\alpha \cdot x_1(k) - (2-\alpha) \cdot x_2(k)}{2}. \tag{3.14}$$

The criterion for the determination of α is to minimize the energy[1] of the difference signal:

$$\mathrm{E}\left\{(y_2(k))^2\right\} \rightarrow \min\,. \tag{3.15}$$

Expanding the energy with Equation 3.14 leads to

$$\mathrm{E}\left\{(y_2(k))^2\right\} = \frac{\alpha^2}{4}\cdot\varphi_{x_1x_1}(0) + \left(\frac{\alpha^2}{2} - \alpha\right)\cdot\varphi_{x_1x_2}(0) + \left(1-\frac{\alpha}{2}\right)^2\cdot\varphi_{x_2x_2}(0) \tag{3.16}$$

[1] Note that the processing is done in a framewise manner so that this is a short term energy that is minimized.

Calculating the derivative with respect to α and setting this equal to zero leads to the following expression for the optimum α:

$$\alpha = \frac{2\varphi_{x_1x_2}(0) + 2\varphi_{x_2x_2}(0)}{\varphi_{x_1x_1}(0) + \varphi_{x_2x_2}(0) + 2\varphi_{x_1x_2}(0)} \,. \tag{3.17}$$

This is indeed always a minimum since the second derivative is always positive apart from the special case that the two signals are identical in amplitude but have different signs which gives a second derivative of zero. However, in this special case, the denominator in Equation 3.17 is zero as well. Hence, this rare case which can only appear in artificial mixtures has to be handled by an exception to ensure stability of the calculation of α.

To illustrate the performance difference between the fixed and the adaptive sum and difference systems, an example simulation was carried out. When calculating the signal energy of $y_2(k)$ with the fixed difference from Equation 3.11 and the adaptive difference from Equation 3.14 for the entire 3GPP dataset [3GP07] consisting of approximately ten minutes of clear and noisy speech from various talkers in different languages as well as music signals, the gain due to the adaptive factor α can be quantified by comparing the energy of the input signals with the energy of the difference signals.

In this evaluation, the processing was done in a blockwise manner with a frame-length of 20 ms. In this dataset, the energy of both channels of the signals is roughly identical:

$$\sum_{k=1}^{N} \left(x_1(k)\right)^2 \approx \sum_{k=1}^{N} \left(x_2(k)\right)^2 \coloneqq \sum_{k=1}^{N} \left(x(k)\right)^2 \,. \tag{3.18}$$

The fixed difference (Equation 3.11) already decreases the energy of $y_2(k)$ in comparison to the input signals. When averaging over all files $p \in P$, the logarithmic energy decrease is

$$\mathrm{E}_p \left\{ 10 \cdot \log_{10} \frac{\sum_{k=1}^{N} \left(x(k)\right)^2}{\sum_{k=1}^{N} \left(y_2(k)\right)^2} \right\} \approx 9\,\mathrm{dB} \tag{3.19}$$

Changing the setup to the adaptive difference (Equation 3.14) improves the performance and leads to an overall energy decrease of 14 dB. Which system, the fixed or the adaptive one, is superior can not be decided in general. This question has to be evaluated for the specific system by finding out if the additional energy decrease is worth the additional data rate for the transmission of the factor α.

3.3 Beamforming

The design of array signal processing systems [HL10] received continuous interest for many applications in the radio frequency domain, e.g., [HLS93] as well as the acoustic domain, e.g., [BW01]. A special form of an array signal processing system in the acoustic domain is the linear microphone array which, due to its physical design, can be integrated easily in many communication systems such as video conferencing terminals. A well designed microphone array is an efficient way to already achieve a decent *Signal-to-Noise Ratio* (SNR) directly at the acoustic frontend if the target signal and the acoustic interferers are spatially separated.

Since this spatial separation is usually given in conferencing scenarios, the use of microphone arrays is especially beneficial. Furthermore, the reverberation as well as the level of diffuse background noise is usually quite low in normal conference rooms. Hence, a microphone array is an efficient device to simultaneously amplify one target speaker while attenuating other speakers and background noise.

When designing and parameterizing microphone arrays, the target is usually to generate a certain reception characteristic. For the far field situation, i.e., at distances from the array that are significantly larger than the physical size of the array setup, there are many known procedures that can be utilized. There are some approaches that are specific for the near field [KAWW96, RG00, DM03, FR11] where the far field designs can only be used to approximately determine the reception characteristic. These approaches however, optimize the reception characteristic only on a (semi-) circular arc at one specific distance from the array. A different design was proposed in [ZGET04] which allows to define a target region in the near field and modify the constraints for an adaptive beamformer accordingly. No approach is known yet that allows to optimize the reception characteristic for an entire area in the near field of the microphone array simultaneously for different distances and angles.

3.3.1 Determination of the Reception Characteristic in the Near Field

The proposed optimization procedure for the weighting coefficients relies on the reception characteristic in the near field of the microphone array. The reception characteristic can be determined in a three-step approach by:

1. simulating or measuring impulse responses $h_{n,m}(k)$ between points in the near field and all microphones,
2. processing these impulse responses with the microphone array to get an overall filter for every point in the near field, and
3. calculating the amplification and attenuation for every point from these overall filters.

Impulse Responses in the Near Field

For the determination of the reception characteristic of the microphone array, impulse responses between positions $\mathbf{p}$ (see Figure 3.4) in the near field of the microphone array and all microphones are necessary. These impulse responses can either be simulated (e.g., by the mirror-image method [AB79]) or measured (cf. 2.2). When using simulated impulse responses, point sources on an appropriately chosen spatial grid (e.g., in a two-dimensional cartesian coordinate system: $\mathbf{p} = (x\ y)^T$) in the near field can be assumed and impulse responses $h_{\mathbf{p}m}(k)$ (with the discrete time index k and the microphone index m) from every point source to every microphone (located at position $\mathbf{p}_m$) in the array can be simulated.

With the impulse responses, the microphone signals $x_m(k)$ can be expressed in terms of filtered versions of the assumed source signal $s(k)$ at position $\mathbf{p}$.

$$x_{\mathbf{p}m}(k) = h_{\mathbf{p}m}(k) * s(k) \tag{3.20}$$

The fact that the optimization procedure works in an identical manner with simulated and measured impulse responses makes it very flexible for different practical application scenarios.

Array Processing

A block diagram of the microphone array can be seen in Fig. 3.4. It consists of a filter-and-sum setup with FIR filters $f_{m,1}(k)$ with L coefficients each at all M microphones. The output $y_{\mathbf{p}}(k)$ of the microphone array depends on the source location $\mathbf{p}$ and can be calculated according to

$$y_{1,\mathbf{p}}(k) = \sum_{m=1}^{\mathrm{M}} f_{m,1}(k) * x_{\mathbf{p}m}(k)\,. \tag{3.21}$$

For every position, the weighted superposition of the individual signals leads

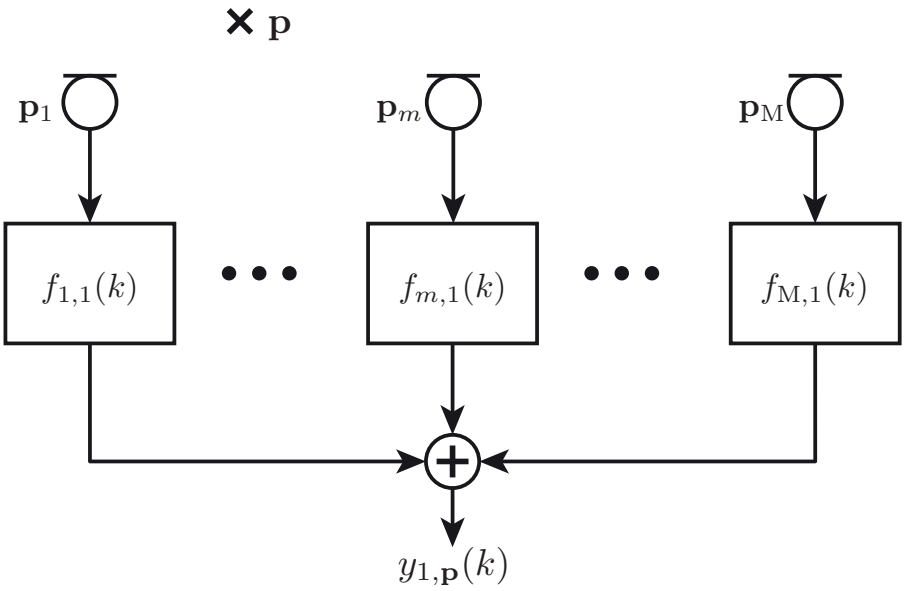


Figure 3.4: Filter-and-sum microphone array

to an effective overall filter $g_{\mathbf{p}}(k)$ since the output signal can be expressed as a filtered version of the source signal.

$$y_{1,\mathbf{p}}(k) = \sum_{m=1}^{\mathrm{M}} f_{m,1}(k) * \Big(h_{\mathbf{p}m}(k) * s(k) \Big) \tag{3.22}$$

Due to the associative property of the convolution, this can be rearranged to

$$y_{1,\mathbf{p}}(k) = \sum_{m=1}^{\mathrm{M}} \Big(f_{m,1}(k) * h_{\mathbf{p}m}(k) \Big) * s(k)\,. \tag{3.23}$$

The overall filter $g_{\mathbf{p}}(k)$ for a source at point $\mathbf{p}$ can hence be determined as

$$g_{\mathbf{p}}(k) = \sum_{m=1}^{\mathrm{M}} f_{m,1}(k) * h_{\mathbf{p}m}(k)\,. \tag{3.24}$$

Calculation of the Reception Characteristic

With the frequency transform of the overall filter $g_{\mathbf{p}}(k)$

$$G_{\mathbf{p}}(f) = \mathcal{F}\left\{ g_{\mathbf{p}}(k) \right\}\,, \tag{3.25}$$

the reception characteristic $S_{\mathbf{p}}(f)$ in dB can be calculated at frequency f for every point $\mathbf{p}$ in the vicinity of the microphone array by

$$S_{\mathbf{p}}(f) = 20 \cdot \log_{10} \left| G_{\mathbf{p}}(f) \right|\,. \tag{3.26}$$

3.3.2 Numerical Optimization

The filter-and-sum microphone array offers $\mathrm{M} \cdot L$ degrees of freedom, the filter coefficients, that have to be set to achieve a certain predefined behaviour of the system. For these filter coefficients, a novel numerical optimization scheme is proposed for the calculation of the filter coefficients in order to mimic a target for the reception characteristic.

Definition of the Target

The target $\hat{S}_{\mathbf{p}}(f)$ for the optimization is defined as a spatial distribution of areas of amplification or attenuation in front of the microphone array. The target speaker shall be in an amplified area $\mathbb{P}_{\text{high}}$ (target value for the reception characteristic S_{high}) while all interferers shall be in attenuated areas $\mathbb{P}_{\text{low}}$ (target value for the reception characteristic S_{low}). This is basically equivalent to defining a target SNR gain between the target speaker and the interferers. The exact location of these areas (and also the target values) is related to the application, e.g., in a conferencing scenario, the target speaker shall be amplified while all interfering sources (such as fans, climate machines or also competing speakers) shall be attenuated. The target can be defined individually for all frequencies but a frequency-independent target is usually appropriate especially for speech communication systems.

$$\hat{S}_{\mathbf{p}}(f) = \hat{S}_{\mathbf{p}} = \begin{cases} S_{\text{high}} & \text{for } \mathbf{p} \in \mathbb{P}_{\text{high}} \\ S_{\text{low}} & \text{for } \mathbf{p} \in \mathbb{P}_{\text{low}} \end{cases} \tag{3.27}$$

An additional advantage of this concept for the determination of the filter coefficients is that the definition of the target areas also allows to include computational complexity considerations within the system design procedure: Larger target areas lead to larger complexity (assuming that the resolution of the spatial grid remains unchanged). The majority of the computational complexity within the optimization process lies in the computation of the error function which will be introduced in the next section. It has to be evaluated only at the points that are in the target area but has to be evaluated frequently within the optimization process.

Error Function and Optimization

The objective of the optimization procedure is to minimize the summed difference Δ_S between the predefined target $\hat{S}$ and the calculated reception characteristic S. The difference is summed over all points for which $\hat{S}_{\mathbf{p}}(f)$ is defined

according to Eq. 3.27 and over all frequencies $f \in [f_{\min} \dots f_{\max}]$ for which the reception characteristic shall be optimized.

$$\Delta_S = \sum_{f=f_{\min}}^{\mathrm{f_{max}}} \sum_{\mathbf{p} \in \left(\mathbb{P}_{\mathrm{high}} \cup \mathbb{P}_{\mathrm{low}}\right)} \left| \hat{S}_{\mathbf{p}}(f) - S_{\mathbf{p}}(f) \right| \tag{3.28}$$

The optimum filter coefficients are determined from this summed level difference in a *Minimum Mean Square Error* (MMSE) sense.

$$\Delta_S^2 \rightarrow \min \tag{3.29}$$

The optimization is carried out by an interior-point algorithm [BGN00] under the constraint that the filter coefficients have to be in the range of $f_{m,1}(k)_{\min} = -1$ and $f_{m,1}(k)_{\max} = 1$. This constraint does not change the shape of the filters, it only limits the maximum amplification that is achievable by the array itself. A subsequent scaling of the output $y(k)$ can be applied to control this.

3.3.3 Performance Example

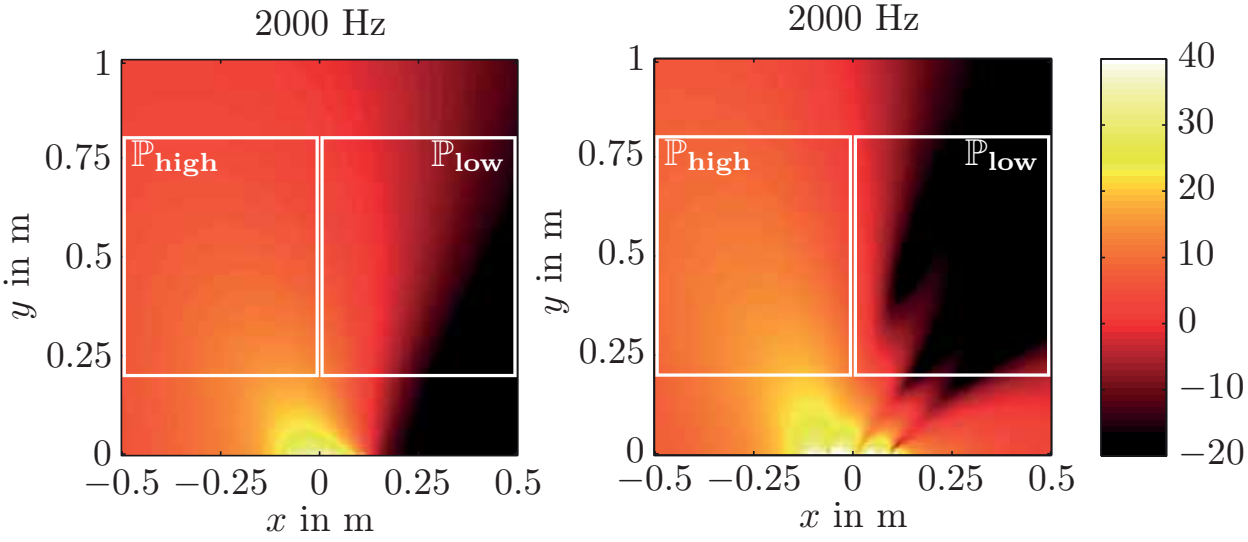
The performance of the novel optimization procedure for the reception characteristic of microphone arrays in the near field is assessed exemplarily by comparing it to the reception characteristic of unoptimized microphone arrays. In a first step, a filter length of $L = 1$ is used in a weighted delay-and-sum setup. After that, the impact of a larger filter length L is quantified as well.

In a possible application, e.g., within a video conferencing system, the simulation of the impulse responses can be fairly simple since conference rooms are usually not highly reverberant. In this case, a simple mirror-image approach or even an approximation by a free field model is suitable. The reception characteristic is visualized here (without loss of generality) for a free field setup since this allows for a clearer evaluation of the impact of the weighting coefficients (an inclusion of real acoustic impulse responses would be straightforward). A comparison of the reception characteristics is given for two different frequencies:

- $f = 2000\,\mathrm{Hz}$ as a medium frequency of the operational frequency range of the microphone array,
- $f = 500\,\mathrm{Hz}$ as a representative for the lower frequencies for which the microphone array should be optimized.

The microphone array is designed to amplify sources on the left ($-0.5\,\mathrm{m} \leq x < 0\,\mathrm{m} \wedge 0.2\,\mathrm{m} < y \leq 0.8\,\mathrm{m}$) while attenuating sources on the right ($0\,\mathrm{m} < x \leq 0.5\,\mathrm{m} \wedge 0.2\,\mathrm{m} < y \leq 0.8\,\mathrm{m}$). For both dimensions ($x$ and y), the density of the spatial grid is set to 0.01 m leading to 3000 points in $\mathbb{P}_{\text{high}}$ and $\mathbb{P}_{\text{low}}$, respectively. These parameters are chosen as a reasonable compromise between precision and computational complexity.

The basis for the comparison in the weighted delay-and-sum setup is the Chebyshev weighting $\mathbf{w}_{\text{Cheb}}$ which is chosen here since it allows to specifiy a minimum attenuation for all side lobes while at the same time also minimizing the width of the main lobe. This combination is very advantageous since it maximizes the SNR between a target area and a diffuse noise field.



Fixed (Chebyshev weighting) Optimized ($L = 1$)

Figure 3.5: Reception characteristics of the microphone array at 2000 Hz

For both the Chebyshev weighting and the optimized weighting, the same delays (which coherently add up the signals from the middle of the target area $\mathbb{P}_{\mathbf{high}}$) are used to allow for a detailed comparison that only takes the effect of the weighting coefficients into account. Additionally, both weightings are parameterized in such a way that they are supposed to achieve a difference in the reception characteristic of 40 dB between the amplified and the attenuated area. For this comparison, a microphone array consisting of 8 sensors with a uniform spacing of 3 cm is used which is centered in the origin of the coordinate system. The sampling frequency for all simulations is $f_s = 48\,\mathrm{kHz}$.

Looking at the performance of the Chebyshev weighting for the 2000 Hz case in the left part of Figure 3.5, there is already a significant level difference between the left and the right side showing that the Chebyshev weighting can be used at

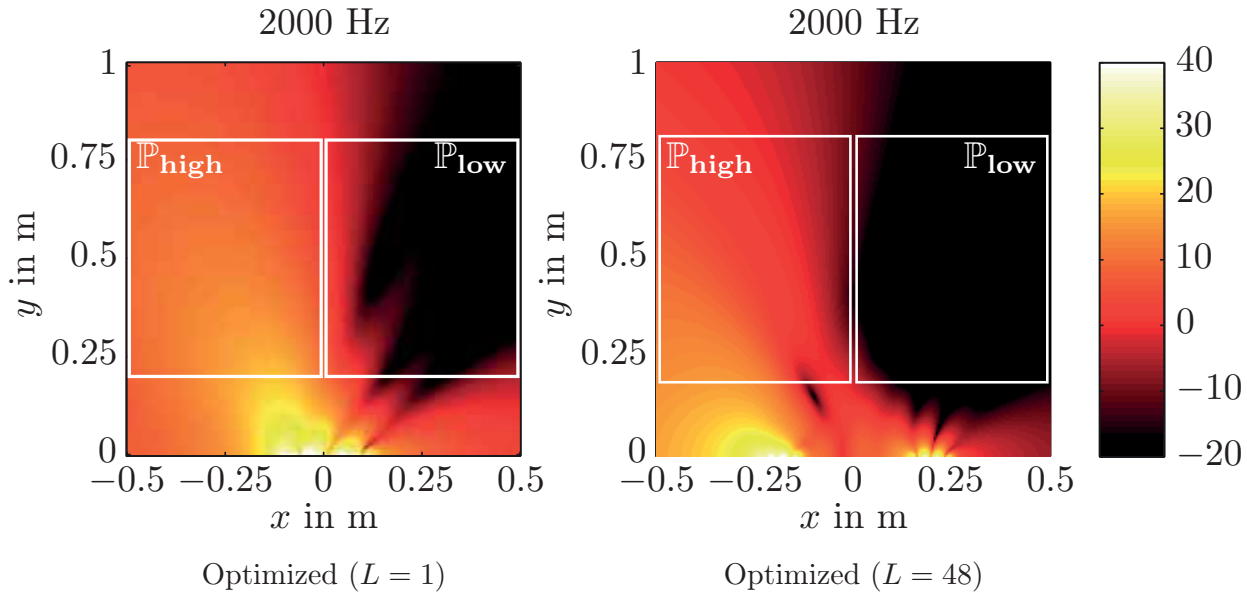


Figure 3.6: Reception characteristics of the microphone array at 2000 Hz for different filter lengths L

this frequency with this microphone array. However, the amplified area clearly extends to the area directly to the right of the center ($0\text{m} < x < 0.25\text{m}$). In contrast with the optimized weighting for just one filter tap, the reception characteristic as depicted in the right part of Figure 3.5 matches the previously defined areas of amplification and attenuation very well. Especially in the transition region around $x = 0\,\text{m}$, a more pronounced border between the amplified and the attenuated area can be observed.

The impact of the longer filter length ($L = 48$ instead of $L = 1$) is visible in Figure 3.6. Both the amplified area and the attenuated area are slightly more homogeneous. However, even for the weighted delay-and-sum setup ($L = 1$), the reception characteristic at this frequency is a good approximation of the target so that the additional gain due to the longer filters is not that significant.

For a frequency of 500 Hz, the reception characteristic of the microphone array with the Chebyshev weighting is depicted in the left part of Figure 3.7. This reception characteristic strongly resembles the one of a single omnidirectional microphone in the origin of the coordinate system. The reception characteristic for the optimized weighting coefficients can be found in the right part of Figure 3.7 where, obviously, some level difference between the left and right side can be observed even for this low operational frequency.

At this frequency, a clear difference between $L = 1$ and $L = 48$ can be observed. While the weighted delay-and-sum setup basically achieves its directivity by placing a null in the direction of the center of the attenuated area, the longer

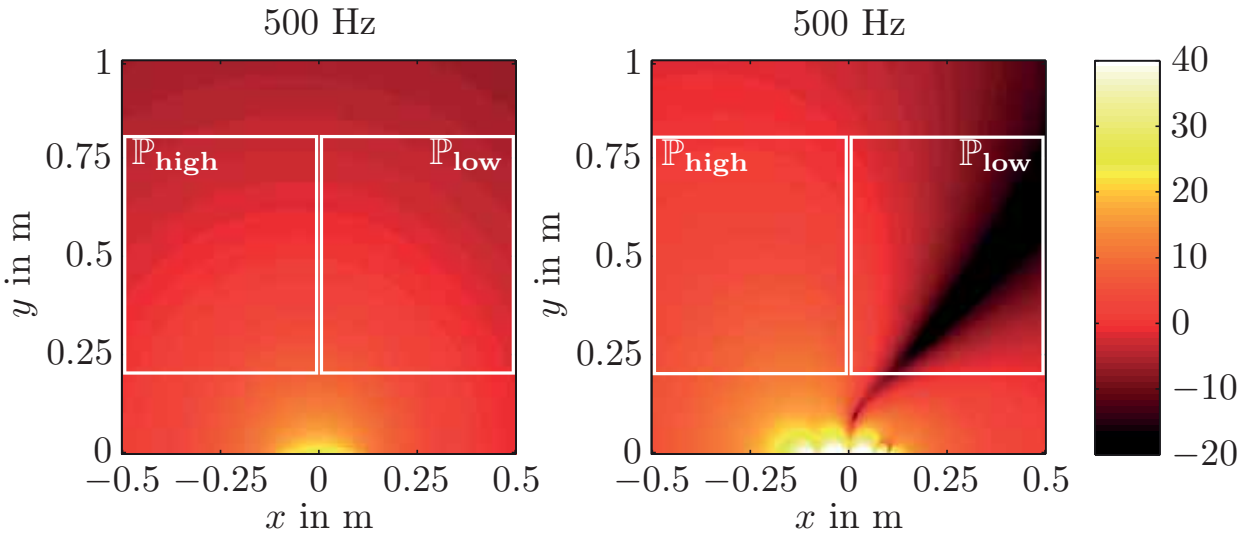


Figure 3.7: Reception characteristics of the microphone array at 500 Hz

filter length allows for a still fairly homogeneous reception characteristic.

The presented beamforming algorithm is hence a very flexible technique that facilitates the approximation of specific targets for the reception characteristic. Depending on the filter length, a good performance even at a low operational frequency can be achieved. This beamforming algorithm has been used to robustly quantify the activity of speakers in a video conferencing scenario [BFS+13].

3.4 Receiver-side Enhancement

So far, the focus was on the transmitting side where different uses for the outer stage were introduced. To conclude this chapter, the receiving side of the outer stage is treated as well which is positioned at the very end of the signal processing chain at the receiving side. The inversion of the different downmixing strategies is not considered here in more detail. The formulas for the inversion of the fixed approaches from Section 3.2.1 as well as the adaptive approaches from Section 3.2.2 are fairly straightforward.

The sum and difference mixing (cf. Equations 3.6 and 3.11) can be inverted by

$$\tilde{x}_1(k) = \tilde{y}_1(k) + \tilde{y}_2(k) \tag{3.30}$$

$$\tilde{x}_2(k) = \tilde{y}_1(k) - \tilde{y}_2(k)\,. \tag{3.31}$$

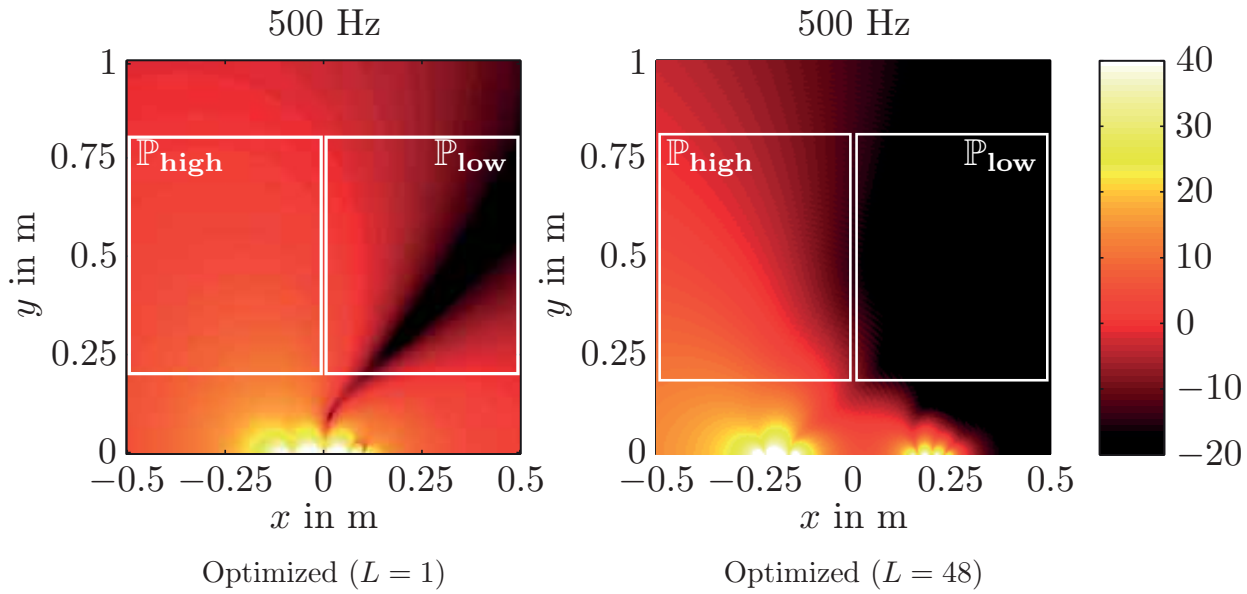


Figure 3.8: Reception characteristics of the microphone array at 500 Hz for different filter lengths L

Similarly, the inversion of the generalized sum and difference (cf. Equations 3.13 and 3.14) can be achieved by

$$\tilde{x}_1(k) = \frac{\tilde{y}_1(k) + \tilde{y}_2(k)}{\alpha} \tag{3.32}$$

$$\tilde{x}_2(k) = \frac{\tilde{y}_1(k) - \tilde{y}_2(k)}{2 - \alpha}\,. \tag{3.33}$$

After this, all the signals are reconstructed and a final stage of signal enhancement can take place. One possibility would be the so-called *Near End Listening Enhancement* (NELE) as presented in, e.g., [SEV06, SV12a, Sau13]. The NELE system delivers a very good system performance with respect to the *Speech Intelligibility Index* (SII). A low complexity alternative based on measurements of the *Speech Transmission Index* (STI) was presented in [SJSV10] and is described and analyzed here. The enhancement procedure is usable for single channel as well as for multi channel transmission systems where particular care has to be taken in order not to alter the spatial cues. A method for signal enhancement in a multi channel environment was proposed in [JSEV10] which can also be the basis for a multi channel application of the enhancement procedure that is presented in the following.

It is known from experience that strong reverberation usually has a detrimental effect on various aspects of speech or audio presentation. Especially speech intelligibility was shown to be severely degraded in reverberant acoustical environments. In [NLT89], it was even shown that the effect of reverberation

on speech intelligibility could not be adequately explained by simple masking effects alone but that a combination of overlap- and self-masking has to be considered.

In contrast to that, it is often argued that having some reverberation can have a positive influence on speech intelligibility [Kut09]. Based on this qualitative argument, the effect of short *Impulse Responses* (IRs) was quantified in [SJSV10] by means of the STI, which is a well developed measure for the intelligibility of speech in various conditions, especially taking into account the effects of additive noise and reverberation. For different scenarios and test signals, different variants of the STI were proposed and extensive testing of the different approaches has been carried out in the past. The so-called *envelope regression method* [LEKP90] was recommended in a recent comparative study [GG04] for the use with speech input signals.

The comparison here is focused on the impact of the chosen impulse response on the speech intelligibility. There are two different types of impulse responses that have to be considered: measured and simulated IRs. There are some measured IRs available covering some environments (from low to high reverberation times) and source-receiver setups (from single to multiple sources and receivers or binaural setups with dummy heads) [JSV09, WGH+06, ADTA01]. For the simulation of IRs, one has the choice of either simulating the entire Room Impulse Response (e.g., by means of the image method [AB79]) or focusing on either the early reflections (e.g., in the form of a sparse IR [BHCN06]) or the diffuse, late reverberation (e.g., by means a statistical model [Pol88]).

The different IRs will be evaluated as enhancement postfilters in an application scenario where speech intelligibility is an absolute necessity: telephony in a mobile, fixed-line, or *Voice over IP* (VoIP) environment. It was shown in [SJSV10] that even codecs that are currently being introduced into the networks fail to reach acceptable STI values especially at lower data rates. One prominent example is the *Adaptive Multi-Rate Wideband* (AMR-WB) codec [ITU03]: its three lowest data rates are not able to provide a good speech intelligibility according to the STI. The approach that is presented here is shown to be capable of consistently improving the intelligibility for this codec.

The remainder of this section is organised as follows: First, the STI in general is shortly introduced in Section 3.4.1 and the specific method that will be used here is presented. A presentation of the different types of IRs follows in Section 3.4.2. Subsequently, the structure of the reverberation-based post-processing is described in Section 3.4.3. Optimized IRs are derived from STI measurements in Section 3.4.4 and explicit recommendations are deduced. Some other possible use cases for the post-processing scheme are presented in

Section 3.4.5.

3.4.1 Speech Transmission Index

The basis for the STI [Int03] was laid in the context of measurements of early very-high-frequency-radio systems. There has been a continuous development in this area for more than three decades now, beginning with the early works of Houtgast and Steeneken [HS71, SH80].

The STI characterizes the system-under-test based on the comparison of two signals: the input (or probe) signal $x(k)$ and the output (or response) signal $y(k)$ with the time index k. The original proposal of measuring STI with an artificial probe signal was later extended by different approaches to use speech as the probe signal. A good overview on the various speech-based STI approaches and a comparison thereof can be found in [GG04]. The basic system that is used for the calculation of the STI in all concepts can be found in Figure 3.9.

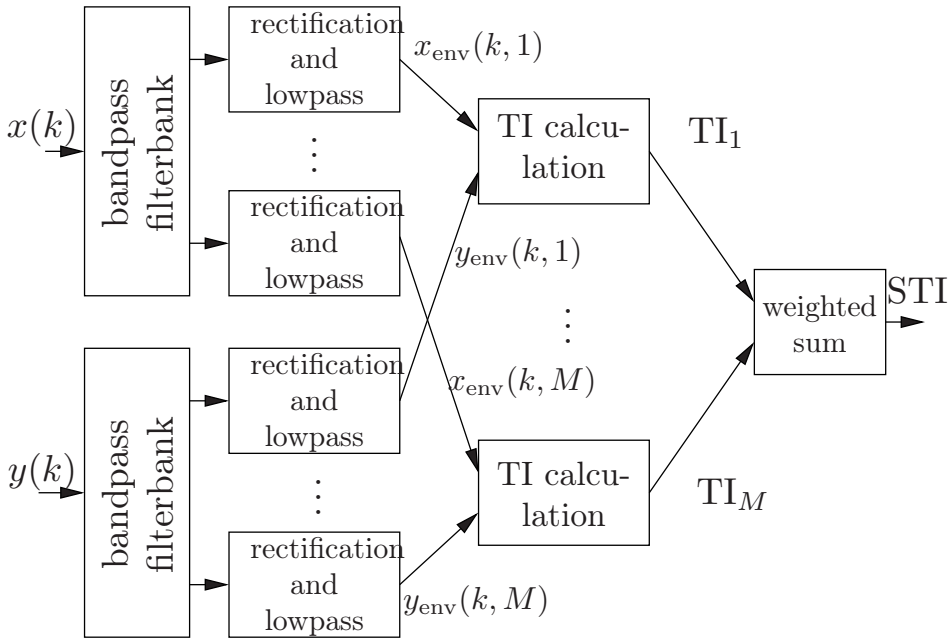


Figure 3.9: Block diagram of the STI calculation.

The STI is calculated as a weighted summation of the individual band transmission indices TI_m. These are calculated in each frequency band $m \in \{1, 2, \ldots M\}$ based on the envelope signals $x_{\mathrm{env}}(k, m)$ and $y_{\mathrm{env}}(k, m)$ of the bandpass-filtered input and output signals $x(k, m)$ and $y(k, m)$.

For the evaluation in this contribution, the so-called *envelope regression method* according to Ludvigsen et al. [LEKP90] is used. The extensive comparison

of the different speech-based STI procedures by Goldsworthy and Greenstein [GG04] has shown that this method leads to equivalent results as the common non-speech-based STI method at a reasonable computational complexity.

The specific property of this method in comparison to other known approaches is that it calculates the apparent Signal-to-Noise Ratio in each band aSNR_m by comparing the input and output envelope signals based on a linear regression analysis. The details can be found in [LEKP90] and [GG04].

3.4.2 Measured and Simulated Room Impulse Responses

When evaluating or developing signal processing algorithms that are related to acoustical reverberation, one has the choice of using either measured or simulated impulse responses. Both approaches have their advantages and disadvantages:

- *Measured impulse responses* inherently capture all properties of real-world environments and are hence more precise when it comes to replicating the reality. On the other hand, there is no infinite number of properly measured IRs available that are representative for all possible application environments. This might lead to overfitting the algorithms to the available datasets.
- *Simulated impulse responses* can be calculated for practically any environment so that there is no risk of developing an algorithm only for a few rooms that happen to be measured in the past. However, simulated impulse responses do not give a perfect representation of every aspect of real IRs.

Real Impulse Responses – the AIR Database

For the evaluations here, impulse responses from the *Aachen Impulse Response* (AIR) database (cf. Section 2.2) are used as measured room impulse responses. The main purpose of this database is the evaluation of speech enhancement algorithms dealing with room reverberation in a binaural application scenario (e.g., hearing aids). For the application as a reverberation postfilter, only single channel IRs can be used. The AIR database contains measurements both with and without a dummy head. For the application in this postfiltering context, the left channel of each measurement without the presence of a dummy head was used.

Depending on the measurement room, the AIR database includes different lengths of the direct path between source and receiver. The details for the excerpt that is used for the evaluation in this contribution can be found in Table 3.1. With this variability, different *Direct-to-Reverberant Energy Ratios* (DRRs) are represented in the excerpt, which allows a first look at the importance of the different parts of the IR for a possible change of speech intelligibility.

Room	Lengths of the direct paths in m
Studio booth	0.5, 1.0 and 1.5
Office room	1.0, 2.0 and 3.0
Meeting room	1.45, 1.7, 1.9, 2.25 and 2.8
Lecture room	2.25, 4.0, 5.56, 7.1, 8.68 and 10.2

Table 3.1: Room configurations for the AIR database.

The room parameters that influence the reverberation characteristics of the measurement rooms differ significantly. While the volume of the studio booth is small (3.00 m × 1.80 m × 2.20 m) and it is specifically designed to have a short reverberation time that is approximately constant over frequency, the lecture room is fairly large (10.80 m × 10.90 m × 3.15 m) and has very reflective surfaces (three walls mostly consist of glass windows, one wall is painted concrete and the floor is parquet). The average reverberation times T_{60} for the four rooms are given in Table 3.2.

Room	Average reverberation time T_{60}
Studio booth	0.12 s
Office room	0.43 s
Meeting room	0.23 s
Lecture room	0.78 s

Table 3.2: Average reverberation times for the different rooms.

Simulation Methods

In addition to the measured impulse responses, two different simulation strategies have also been tested with respect to their applicability for improving speech intelligibility. Two significantly different models were chosen due to the fact that real-world impulse responses can be divided into two parts:

- early reflections (including the direct path) and
- late, diffuse reverberation.

In order to separately examine the influence of both components of the IR, one of the models only simulates the late reverberant tail while the other one only consists of a few strong early reflections.

The representative for the late reverberant tail is the design according to the exponential decay model by Polack [Pol88]. It is trying to mimic the late reverberation properties of real environments (e.g. [Kut09]) by shaping the envelope of white noise.

In the first step, this model generates a white Gaussian noise signal $n(k)$ of length $T \cdot f_s$ with the target duration T of the impulse response and the sampling frequency f_s. This signal has zero mean and is uncorrelated.

$$E\{n(k)\} = 0 \tag{3.34}$$

$$E\{n(k) \cdot n(k+\kappa)\} = 0 \quad \text{for} \quad \kappa \neq 0 \tag{3.35}$$

This noise $n(k)$ is then shaped by an exponential decay $b(k)$ which has the same length $T \cdot f_s$ as the noise and can be parameterized by the reverberation time T_{60}:

$$b(k) = \mathrm{e}^{-\frac{3 \cdot \ln(10)}{T_{60}} \cdot k}. \tag{3.36}$$

The final impulse response $h(k)$ can then be calculated as the multiplication of the two signals:

$$h(k) = n(k) \cdot b(k)\,. \tag{3.37}$$

The model can be extended to include a delay for representing the length of the direct path. For the application as a signal processing postfilter, this is omitted as it would only cause additional processing delay which is generally undesirable.

This model does not consider early, individual reflections, which for most rooms form the first 50-80 ms of the IR after the arrival of the sound on the direct path. Instead, it focuses on the diffuse reflections that occur later in the IR.

An alternative that emphasizes the strong individual components that are present in real-world acoustic environments are sparse impulse responses.

These consist of just very few components $h(k) \neq 0$. In the most simple setup, such an IR only consists of two coefficients: the direct path at $k = 0$ with the amplitude h_{direct} and a single reflection at $k = k_1$ with the amplitude $h_{\text{reflection}}$.

It can be expressed by a two-tap FIR filter with transfer function

$$H(z) = h_{\text{direct}} + h_{\text{reflection}} \cdot z^{-k_1} \,. \tag{3.38}$$

Just like in the case of the Polack model, a delay for the length of the direct path is not included.

3.4.3 Postfilter Design

The structure of the system that is necessary for investigating the properties of the measured or simulated room impulse responses is depicted in Fig. 3.10. It consists of an FIR filter which is used for post-processing of the respective system (e.g., speech codec).

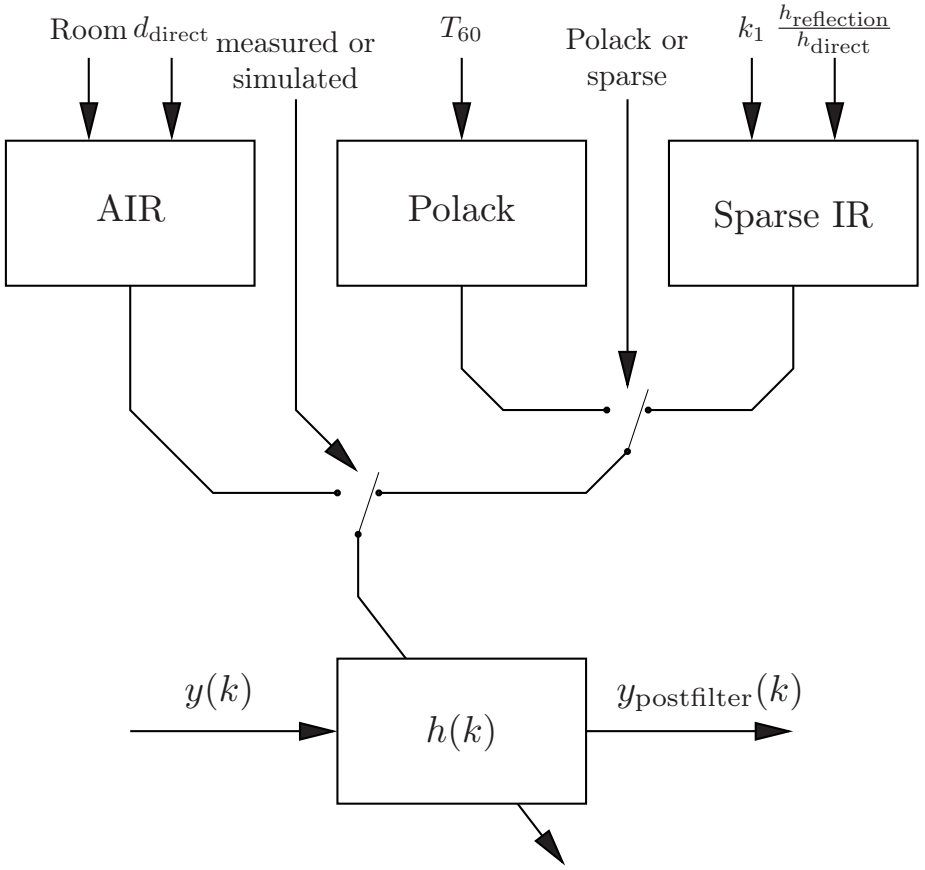


Figure 3.10: Enhancement postfilter and candidate room impulse response models.

There are various parameters that can be set depending on the type of impulse response that is used. For the measured IRs, one has the choice between four

different rooms with three to six different lengths of the direct path between source and receiver.

For the simulated IRs, the first choice has to be between the two models: either the statistical model from Polack or the sparse impulse response. The statistical model can then be parameterized by the reverberation time T_{60}. The sparse IR needs two input parameters: the position k_1 of the second filter tap in relation to the first tap and the amplitude relation $\frac{h_{\text{reflection}}}{h_{\text{direct}}}$ between the two filter taps.

To allow for a fair comparison between the different IRs, an amplitude normalization of the impulse response is carried out. This ensures that the STI is unaffected by possibly different energy levels of the signals. This is also the reason why the amplitude relation is a sufficient description of the sparse IR.

As described in the last section, the two simulation models do not incorporate an additional delay for the length of the direct path so that they inherently do not cause an additional algorithmic delay. The measured impulse responses do have a fundamental delay t_{direct} that is related to the length of the direct path d_{direct} by

$$t_{\text{direct}} = c_0 \cdot d_{\text{direct}} \tag{3.39}$$

with c_0 as the speed of sound. Removing the first $t_{\text{direct}} \cdot f_s$ samples from the impulse response is a simple yet effective countermeasure and leads to an identical fundamental delay of zero samples for all IRs. It is important to note that this does not make the different measured IRs from one room identical as they still exhibit, e.g., different DRRs.

The complexity of the postfilter is directly proportional to the number of non-zero filter taps. Each non-zero filter tap requires one multiply and one add operation per sample. Since this can be computationally expensive for long filters if the processing is carried out in the time domain, frequency domain processing could be used in those cases to increase the efficiency. Postfiltering with the sparse IR on the other hand can easily be executed in the time domain due to the very low number of non-zero taps.

3.4.4 Measurements

The proposed post-processing was evaluated as an enhancement for the AMR-WB speech codec [ITU03]. The evaluation is based on single channel transmission – the application of the enhancement structure in a multi channel

transmission system is straightforward: Identical impulse responses have to be used in all channels (cf. the system from [JSEV10]). The NTT speech corpus [NA94] was used as the dataset for the evaluation.

As a reference, the STI was calculated between the clear speech signal as the probe signal and the output of the AMR-WB speech codec (encoding and decoding without transmission errors) as the response signal. Each file in the speech corpus was processed individually and the STI values were averaged, the resulting mean values are given in Table 3.3. Usually, systems with an STI of 0.6 or greater are considered good [HSA+02] while a value of 0.5 should at least be reached for an acceptable intelligibility.

Data rate in kbit/s	Average STI
6.60	0.4693
8.85	0.5416
12.65	0.5983
14.25	0.6118
15.85	0.6242
18.25	0.6436
19.85	0.6494
23.05	0.6686
23.85	0.6703

Table 3.3: Average STI values for the different possible data rates of AMR-WB.

It can be seen from this evaluation that the speech intelligibility of the transmission with the AMR-WB speech codec does not achieve a good intelligibility according to the STI and that the lowest data rate does not even reach acceptable intelligibility.

The first measurement results for the post-processing scheme are those with measured impulse responses from the AIR database in four different rooms, they can be found in Figure 3.11.

Since the AMR-WB speech codec operates at a sampling frequency of $f_s = 16\,\text{kHz}$, a downsampled version of the AIR database was used. The dotted line marks the average STI for the particular data rate of the AMR-WB speech codec without post-processing. It can be seen that most impulse responses decrease the STI with the notable exception of very short lengths of the direct path in the less reverberant rooms (studio booth and meeting room), where an increase in STI for the lower data rates can be observed.

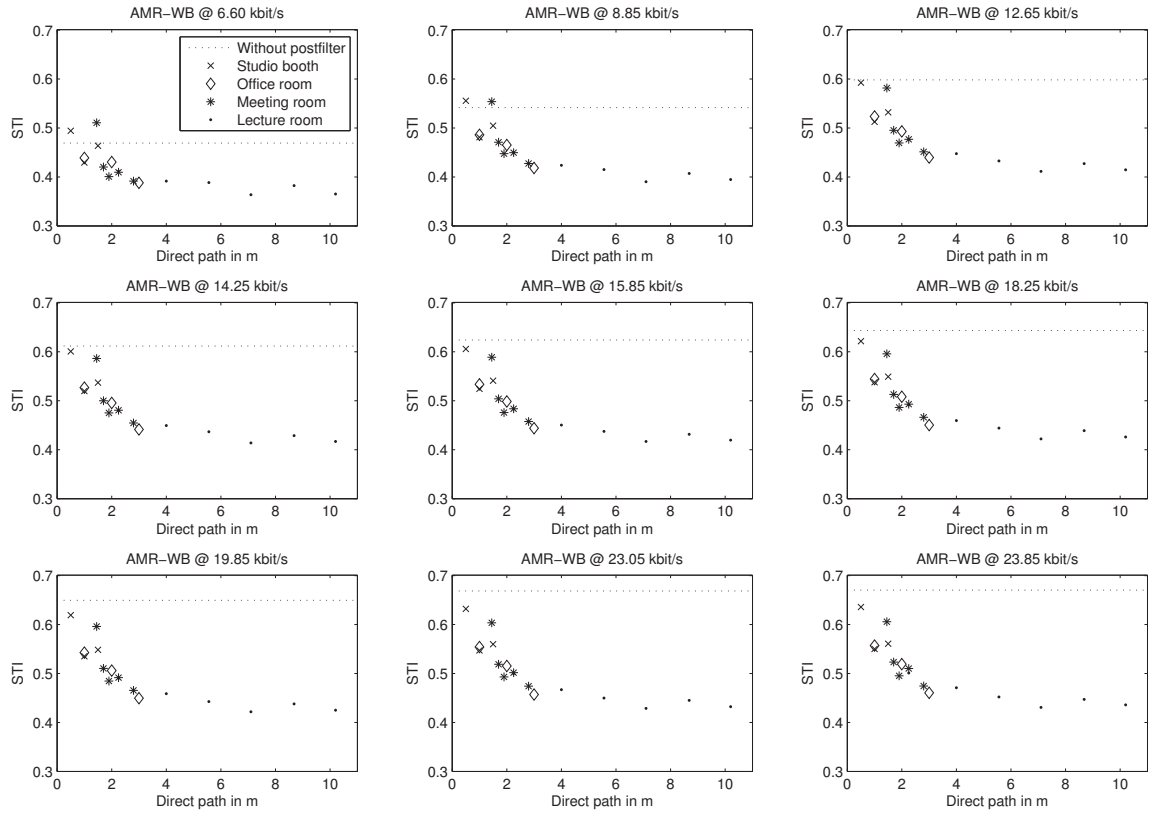


Figure 3.11: STI for AMR-WB after postfiltering with impulse responses from the AIR database.

The resulting STI values for the nine different operation modes of AMR-WB in combination with the proposed postfilter for the model of Polack are depicted in Figure 3.12. Again, the dotted line marks the average STI without post-processing. It can be seen that even for low data rates and very short reverberation times T_{60}, there is no increase in STI and especially for higher data rates, a significant drop in STI is obvious.

The last results are those for a postfiltering with the sparse IRs with just two non-zero coefficients in $h(k)$, which can be found in Figure 3.13. For all data rates, the largest STI values can be observed for the case that the second non-zero coefficient directly follows the direct path (i.e., $k_1 = 1$). The behaviour with respect to the amplitude relation $\gamma = \frac{h_{\text{reflection}}}{h_{\text{direct}}}$ is less explicit, the changes between the values are significantly smaller. For the two lowest data rates, the maximum STI can be found for $\gamma = 1$ while for all the other data rates, a quotient of $\gamma = 0.3$ leads to the largest STI. An overview on the achievable STI in comparison to the STI without post-processing can be found in Table 3.4.

The STI is known to be well-correlated to the intelligibility of reverberant speech [GG04, HSA+02]. Informal listening tests support the increase in intel-

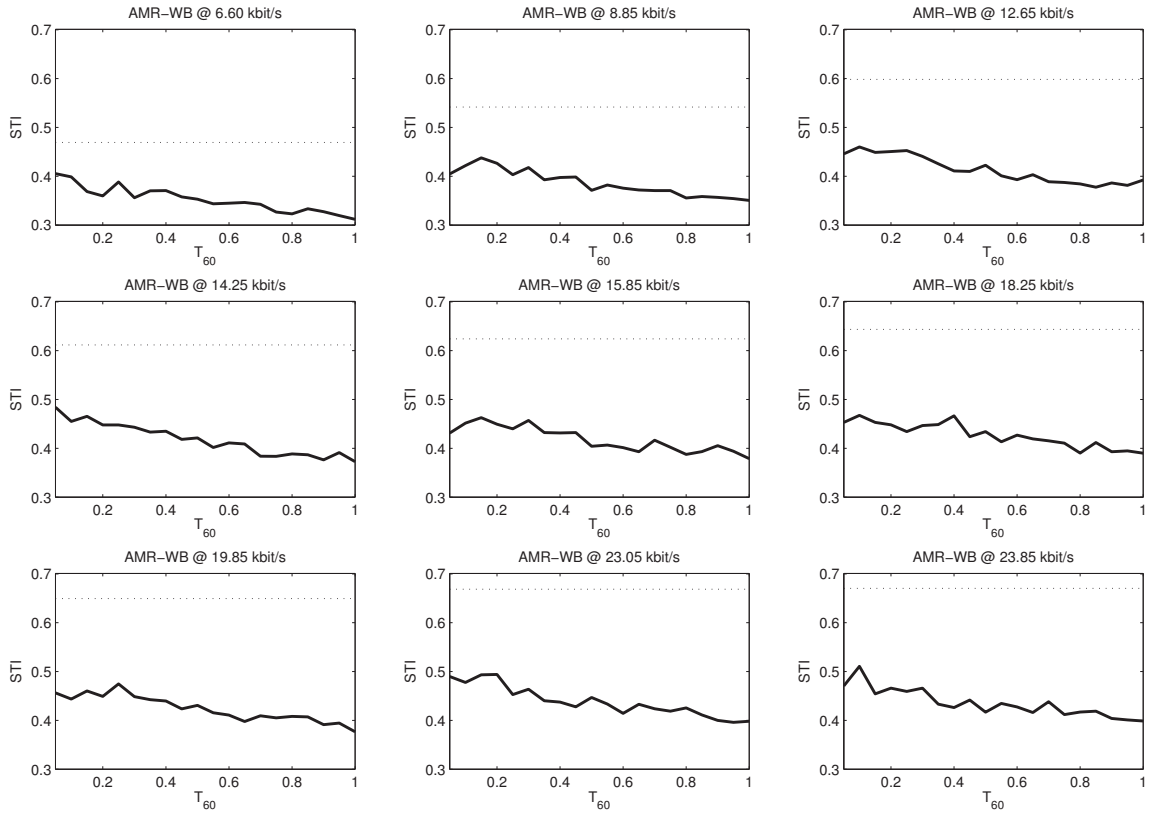


Figure 3.12: STI for AMR-WB after postfiltering with impulse responses according to the model of Polack.

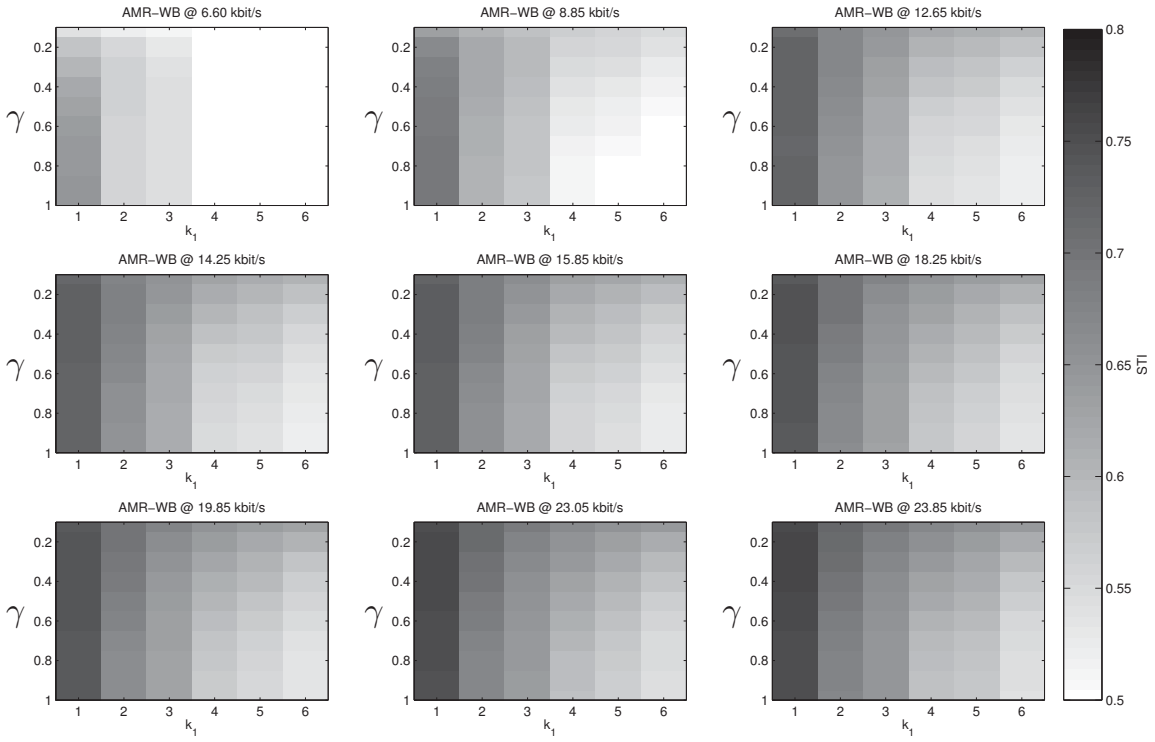


Figure 3.13: STI for AMR-WB after postfiltering with sparse IRs.

Data rate in kbit/s	STI without postfiltering	Achievable STI
6.60	0.4693	0.6445
8.85	0.5416	0.7004
12.65	0.5983	0.7376
14.25	0.6118	0.7456
15.85	0.6242	0.7533
18.25	0.6436	0.7660
19.85	0.6494	0.7691
23.05	0.6686	0.7815
23.85	0.6703	0.7831

Table 3.4: Average STI values for the different possible data rates of AMR-WB and the maximum STI values for postfiltering with sparse IRs.

ligibility that is indicated by the STI for the sparse IRs.

3.4.5 Results

The strong individual reflections that are present in the first part of natural room impulse responses are said to have a positive effect on speech intelligibility. A quantitative study of the effect was carried out based on the *Speech Transmission Index* (STI), a well-developed measure for speech intelligibility in various adverse scenarios. An enhancement postfilter was devised which synthetically adds reverberation based on different types of impulse responses. This postfilter was implemented for the AMR-WB speech codec in order to explicitly determine which part of the impulse response could cause a reproducible and significant increase in STI.

Measured room impulse responses were shown to increase the STI for short lengths of the direct path in smaller rooms and only at very low data rates. In contrast to that, a clear decrease in STI could be observed for bigger rooms and bigger lengths of the direct path (i.e, smaller DRRs).

Postfiltering with simulated IRs leads to ambiguous results. Impulse responses that were designed according to the model of Polack and thus mimic the late reverberant properties do not offer any gain in STI. The sparse IRs on the other hand can be parameterized to significantly increase the STI even for the highest data rates of the AMR-WB speech codec. Optimum amplitude relations between the two taps of the impulse response could be derived that depend on the on the operation mode of AMR-WB.

Reverberation-based post-processing could also be applied for speech enhancement techniques. A small amount of artificial reverberation could help to conceal signal processing artifacts. Additionally, the positive effect of a certain amount of reverberation on the perceived audio quality is well known from the recording of music performances. This so-called comfort reverb leads to small temporal smearing of the speech or audio material which also overshadows, e.g., small intonation errors. Due to this and in view of the ongoing convergence of speech and audio coding, the proposed reverberation postfiltering might also be used to facilitate a better transmission of music with state-of-the-art speech codecs. Possible use-cases for this could include improving the perceived quality for music during regular phone calls as well as streaming applications.

3.5 Conclusion

Three different possible applications of the outer stage were presented in this chapter. After a short review of mixing strategies that will be utilized later on in the signal transmission part of this work, a concept for the design of beamforming algorithms was presented. It consists of a numerical optimization scheme for the filter coefficients of a filter-and-sum array. The optimization scheme allows to optimize the entire reception characteristic in the vicinity of a microphone array at once. The reception characteristic with optimized filter coefficients was shown to match the target characteristic very well.

In a practical application, e.g., within a conferencing system, the optimization scheme is advantageous as it can be used very flexibly due to the fact that it works with simulated as well as measures impulse responses, can be parameterized for lower complexity, and does not rely on any specific microphone array geometry.

Finally, a scheme was presented that can be utilized on the receiving side to increase the speech intelligibility for coded transmissions. The scheme exploits knowledge from the field of room acoustics on the impact of reverberation on the intelligibility of speech. A quantitative study was carried out which leads to an optimized postfilter design that was shown to robustly increase the STI.

4

Inner Stage – Predictive Multi Channel Coding

After the outer stage (see Chapter 3), the signals $y_1(k) \dots y_{\mathrm{C}}(k)$ are processed in the inner stage (depicted in Figure 4.1) which is described in this chapter. On the transmitting side, a predictive multi channel encoding is the final step in preparing the signals for transmission. The hierarchical concept is constructed around the transmission of one main channel and a set of prediction error signals. The receiving side of the inner stage then decodes these signals and reconstructs the input signals.

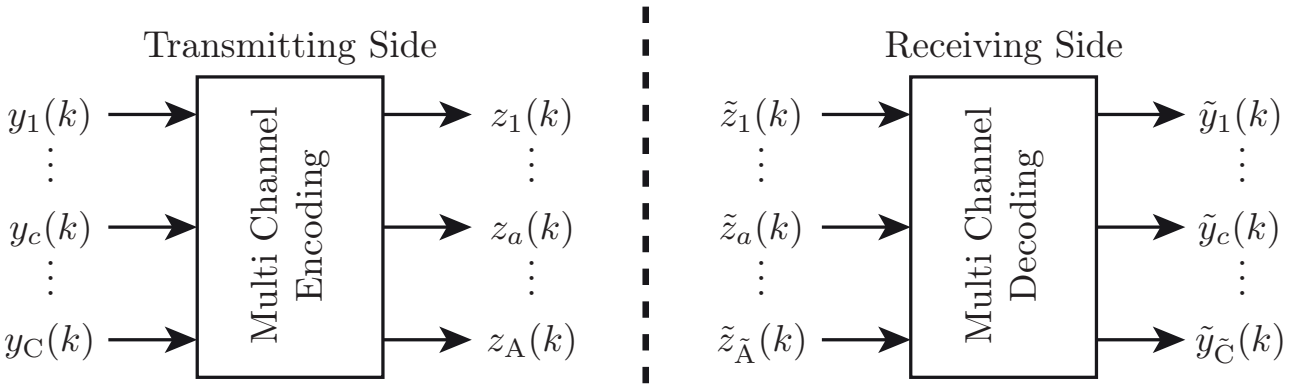


Figure 4.1: Inner stage of the multi channel signal processing system

The hierarchical coding structure is first presented and its properties are explained. The coding scheme is based on a single channel downmixing process (done in the outer stage, cf. Sections 3.2.1 and 3.2.2) followed by predictions of the input signals. Symmetries in the prediction filter coefficients and the prediction errors allow to reduce the number of channels which need to be transmitted.

Subsequently, a practical system is derived that can be combined with standardized codecs for the transmission of the main channel and the prediction errors. As an alternative and for an evaluation of the performance of the hierarchical prediction step alone, a combination with logarithmic quantizers for main channel and prediction errors is considered as well. A detailed evaluation with respect to the achievable prediction gain and the quantizer setup allows to correctly parametrize the encoding system.

Parts of the predictive structure were presented in [KV08, SKV09, SV12b], detailed analyses of the system under different circumstances are presented here along with a novel approach for noise shaping in multi channel predictive systems and complexity considerations especially focusing on the decoding complexity.

It is explained how the proposed system can be combined with existing single channel communication systems in a hierarchical manner. This is particularly attractive since the system introduces only an almost negligible additional algorithmic delay and its audio quality scales very well with the available data rate.

In the following, the predictive structure is introduced in Section 4.1 where a particular focus is put on the decoding complexity. The determination of the optimal filter coefficients for this setup is treated in Section 4.2 before concepts for the inclusion of noise shaping into this system are developed in Section 4.3. The performance of the different aspects of the system is evaluated in Section 4.4 where important system parameters are set based on the results of the evaluation. An application example is presented in Section 4.5 which combines the presented approach with a standardized codec to form a usable system for the transmission of multi channel signals. Concluding remarks on the transmission system are finally given in Section 4.6.

4.1 Structure of the Hierarchical Coding Scheme

The hierarchical coding scheme constitutes the inner stage from Figures 1.2 and 1.3. On the encoding side, it has C input channels and A output channels.

The basic concept is an inter-channel prediction scheme utilizing a specific case of the downmixing stage (cf. Section 3.2) and the original input signals. The downmix signal $z_{\mathrm{A}}(k) = y_{\mathrm{C}}(k)$ is used as the basis for the predictions which are carried out as depicted in Figure 4.2.

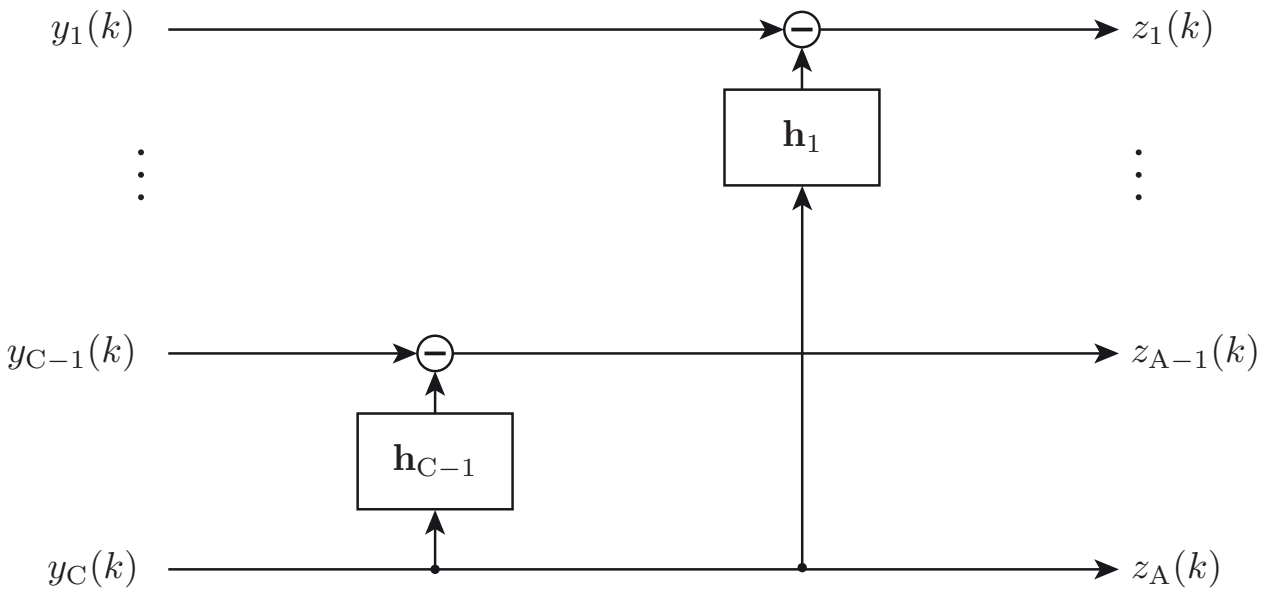


Figure 4.2: Basic Encoding Structure of the Multi Channel Coding System

The downmix or main channel signal $z_A(k)$ that is used in the hierarchical concept presented in this chapter is the normalized sum from Equation 3.7 which is a core component of the transmission as certain advantageous properties of the system can be derived from this downmixing variant.

$$z_A(k) = y_C(k) = \frac{\sum_{m=1}^{M} x_m(k)}{M} \tag{4.1}$$

In a heterogeneous network of multi channel and single channel devices, this main channel signal can be directly used by the single channel devices without the need for additional processing. This makes the proposed structure very attractive for a transition period when new multi channel devices have entered the market while a significant number of single channel devices is still being used. This structure shares some similarities with the matrix decoding techniques that can be found, e.g., in Dolby Pro Logic and Dolby Digital Surround EX. However, the proposed predictive coding scheme is significantly more flexible due to the filter coefficients which are adapted to the current signal.

Additionally, all the original input channels $x_1(k) \dots x_M(k)$ are passed on in a delayed form by the mixing stage according to

$$y_i(k) = x_i(k - \tau) \quad \forall \quad i \leq M\,. \tag{4.2}$$

This is equivalent to setting $C = M + 1$ and choosing the downmixing filters

$f_{m,c}(k)$ to be

$$f_{m,c}(k) = \begin{cases} \frac{1}{\mathrm{M}} \cdot \delta(k) & c = \mathrm{C} \\ \delta(k-\tau) & c = m \\ 0 & \text{else} \end{cases} \tag{4.3}$$

With the unconstrained *Finite Impulse Response* (FIR) filters $\mathbf{h}_1 \ldots \mathbf{h}_{\mathrm{C}-1}$, both amplitude and phase relations between the downmix and the individual input channels can be fully utilized in the prediction process. The prediction errors $z_a(k)$ are then calculated as

$$z_a(k) = y_c(k) - \sum_{\lambda=0}^{L-1} h_a(\lambda) \cdot z_{\mathrm{A}}\,(k-\lambda)\ . \tag{4.4}$$

The delay τ is inherently determined by the degree of the prediction filters: The filter length amounts to $L = 2 \cdot \tau + 1$. This choice for the filter length leads to a symmetric structure: The delayed sample $y_c(k-\tau)$ in the input channel is predicted from the sample in the main channel with identical delay $z_{\mathrm{A}}\,(k-\tau)$ and from both τ newer and τ older samples. Note that this does not imply that the filter itself is symmetric (i.e., linear phase). The filter coefficients for one channel a can be collected in a vector $\mathbf{h}_a$:

$$\mathbf{h}_a = \Big(h_a(0) \ldots h_a(L-1) \Big)^T \tag{4.5}$$

The number A of sent signals is equal to the number C of signals after the mixing stage which itself is one larger than the number of input signals M. For now, this increases the number of signals that need to be transmitted. The decoding structure used to reconstruct the input signals $x_1(k) \ldots x_{\mathrm{M}}(k)$ is depicted in Figure 4.3. When assuming error free transmission (i.e., $\tilde{z}_a(k) = z_a(k)$ and $\tilde{h}_a(k) = h_a(k)$), the exact inversion of the encoding process is possible. It is important to mention that the decoding process does not require inverse filters since the basis for the prediction $z_{\mathrm{A}}(k)$ (resp. $\tilde{z}_a(k)$) is available both at the encoder and the decoder. The inversion of the encoding process can then be done by rearranging Equation 4.4 to

$$\tilde{y}_c(k) = \tilde{z}_a(k) + \sum_{\lambda=0}^{L-1} \tilde{h}_a(\lambda) \cdot \tilde{z}_{\mathrm{A}}(k-\lambda)\,. \tag{4.6}$$

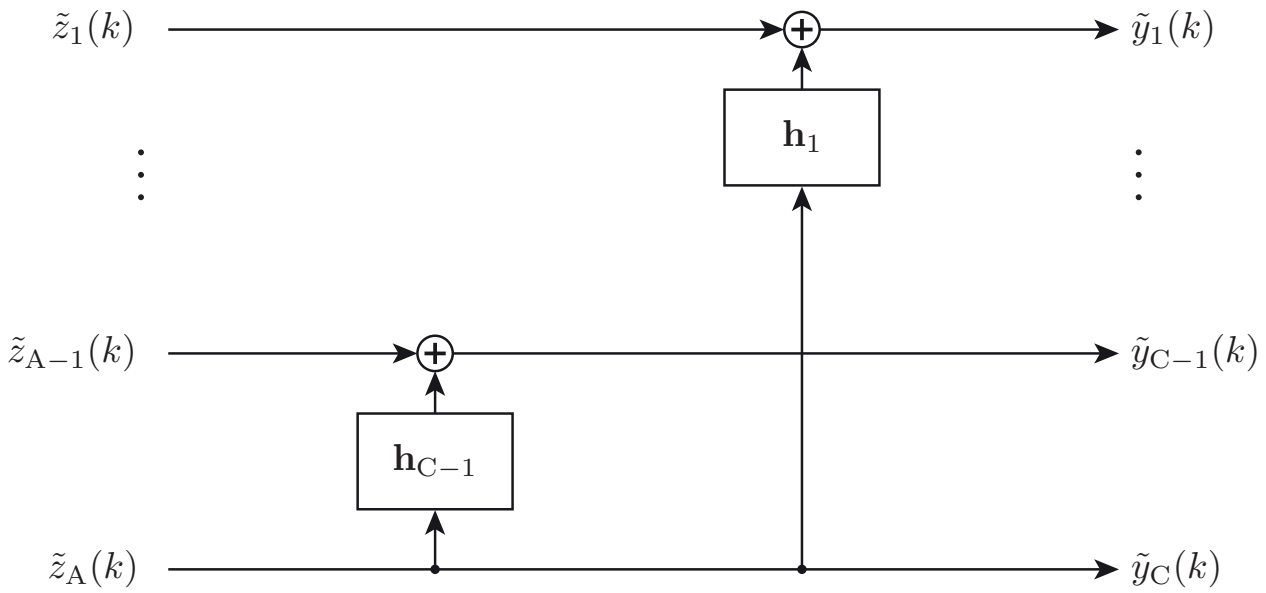


Figure 4.3: Basic Decoding Structure of the Multi Channel Coding System

This decoder setup works flawlessly but there are certain properties of the system that can be exploited to decrease the data rate for the transmission as well as the computational complexity. Without introducing the entire procedure for the determination of the filter coefficients, some of the findings from Section 4.2 are already stated here. It was shown in [KV08] for a related system setup that when calculating the filter coefficients $\mathbf{h}_a$ according to Equation 4.20 and the prediction error signals $z_a(k)$ for all $\mathrm{A}-1$ channels, some advantageous symmetries can be found (the derivations for these symmetries can be found in Appendix B)

- The sum of all vectors of filter coefficients equals a vector with only zeros and a single one in the middle:

$$\sum_{a=1}^{\mathrm{A}-1} \mathbf{h}_a = \Big(\underbrace{0 \ldots 0}_{\tau} \quad 1 \quad \underbrace{0 \ldots 0}_{\tau} \Big)^{\mathrm{T}} \tag{4.7}$$

- The sum of all prediction errors equals zero at all times:

$$\sum_{a=1}^{\mathrm{A}-1} z_a(k) = 0 \tag{4.8}$$

Based on these findings, the sum signal and only $\mathrm{A}-2$ prediction errors and $\mathrm{A}-2$ sets of filter coefficients have to be calculated and transmitted to the decoder as the missing values can be reconstructed by applying Equations (4.7)

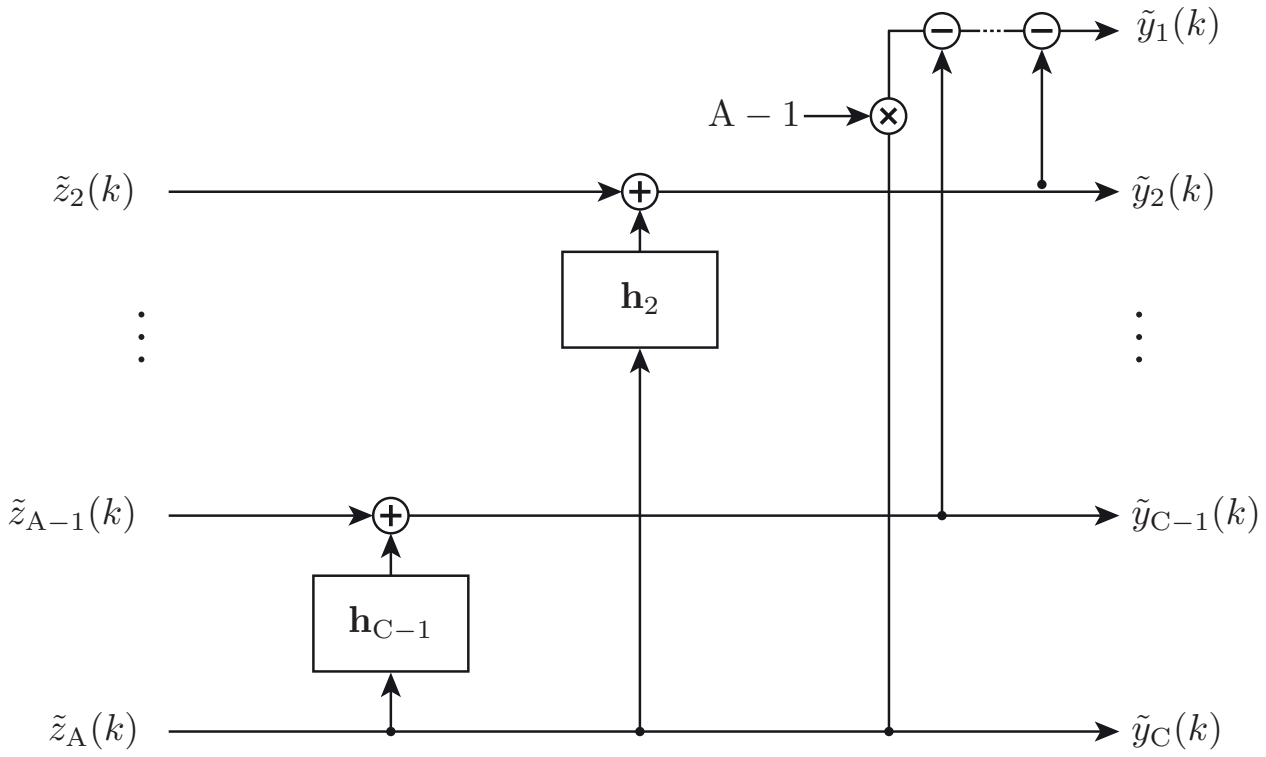


Figure 4.4: Modified Decoding Structure of the Multi Channel Coding System

and (4.8). This can be done in an efficient way by the decoding structure which is depicted in Figure 4.4.

Therein, the reconstruction of all output channels but one is achieved by an reversion of the prediction process from the encoder:

$$\tilde{y}_a(k-\tau) = \tilde{z}_a(k) + \sum_{\lambda=0}^{L-1} h_a(\lambda) \cdot \tilde{z}_{\mathrm{A}}(k-\lambda)\,. \tag{4.9}$$

The final output channel, e.g., signal $\tilde{y}_1(k)$ (w.l.o.g) is then calculated using

$$\tilde{y}_1(k-\tau) = (\mathrm{A}-1) \cdot \tilde{z}_{\mathrm{A}}(k-\tau) - \sum_{a=2}^{\mathrm{A}-1} \tilde{y}_a(k-\tau)\,. \tag{4.10}$$

It can directly be seen that this structure allows for a perfect reconstruction of the input signals given that the transmission introduces no errors (i.e., $\tilde{z}_a(k) = z_a(k) \,\forall\, a$ and $\tilde{\mathbf{h}}_\mathbf{a} = \mathbf{h}_\mathbf{a} \,\forall\, a$).

An overview of the decoding complexity of the new structure in comparison to the unmodified structure from Figure 4.3 is given in Table 4.1.

The modified decoder requires fewer multiplications as long as there are any filters $\mathbf{h}_a$ in the system, i.e., $L > 0$, and it also requires fewer additions as long

	Basic decoder (Figure 4.3)	Modified decoder (Figure 4.4)
Multiplications	$(\mathrm{A}-1)\cdot L$	$(\mathrm{A}-2)\cdot L+1$
Additions	$\mathrm{A}\cdot L$	$(\mathrm{A}-2)\cdot(L+1)$

Table 4.1: Complexity of the previous and new decoding structure

as the filters are reasonably long in comparison to the number of channels, i.e., $L > \frac{\mathrm{A}-2}{2}$.

Hence, especially for the transmission of stereo signals ($\mathrm{A} = 3$, currently probably the most important practical use-case), the computational complexity is significantly reduced.

When looking at the special case of stereo signals, it is worth comparing this structure to the well known mid side stereo coding scheme that is depicted in a full band variant in Figure 4.5 but can also be applied to subband signals produced by a suitable filterbank [JF92].

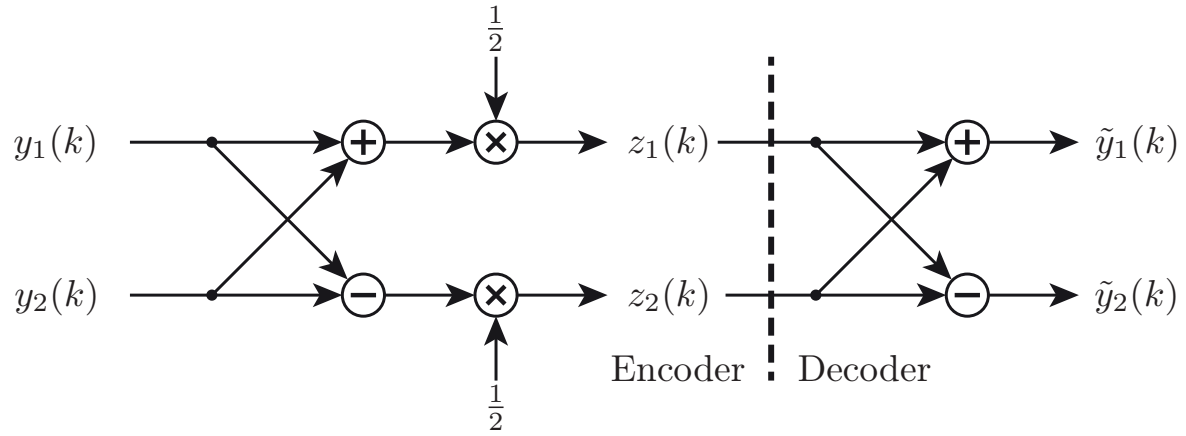


Figure 4.5: Known fullband setup for mid side stereo coding as a special case of the presented predictive coding scheme

In this scheme, the mid channel signal $z_1(k)$ and the side channel signal $z_2(k)$ are calculated from the left audio channel $y_1(k)$ and the right audio channel $y_2(k)$ in analogy to Section 3.2.1 by

$$z_1(k) = \frac{y_1(k) + y_2(k)}{2} \tag{4.11}$$

$$z_2(k) = \frac{y_1(k) - y_2(k)}{2}\,. \tag{4.12}$$

The normalization by a factor of $\frac{1}{2}$ is not necessary for the idea of mid side stereo coding in principle but it is necessary at some stage of the transmission system to ensure that the amplitudes of the output signals are not larger than the amplitudes of the input signals.

Both signals are transmitted to the decoder by means of possibly different single channel transmission systems, e.g., codecs like the ITU G.726 [ITU90]. At the receiving side, the left and right channel signals are reconstructed from the received versions of the mid channel signal $\tilde{z}_1(k)$ and the side channel signal $\tilde{z}_2(k)$ (in Figure 4.5, the transmission is assumed to be perfect, i.e., $\tilde{z}_i(k) = z_i(k)$) as

$$\tilde{y}_1(k) = \tilde{z}_1(k) + \tilde{z}_2(k) \tag{4.13}$$

$$\tilde{y}_2(k) = \tilde{z}_1(k) - \tilde{z}_2(k) \tag{4.14}$$

The rationale behind this structure is that in many audio signals, a strong mid channel signal $z_1(k)$ will result so that the signal variance of $z_1(k)$ is significantly higher than that of $z_2(k)$ which can be exploited to reduce the data rate that is necessary for the transmission of the signals. In addition, the mid channel signal can also be used as a backwards compatible signal between mono and stereo transmission with FM radio transmission as one prominent example.

The two channel version of the presented structure (i.e., C = A = 2 in Figures 4.2 and 4.4) can be seen as a generalization of the mid side stereo coding scheme. The main channel signal $z_{\mathrm{A}}(k)$ from Equation 4.1 is identical to the sum signal $z_1(k)$ from Equation 4.11. The second channel, the difference signal $z_2(k)$ from Equation 4.12 results for the presented approach with the following parameters:

$$\tau = 0 \tag{4.15}$$

$$h_2(0) = 0.5\,. \tag{4.16}$$

It is shown in the next section how, in the more general case, these filter coefficients are determined and later on that the additional degrees of freedom that are available in the new structure provide a significantly better performance than the fixed systems.

4.2 Optimal Filter Coefficients

While the new structure of Figures 4.2 and 4.4 has different degrees of freedom where it can be parameterized, the main adaptivity during operation lies in the filter coefficients. The other aspects like, e.g., number of channels, filter length, quantizer types or quantizer word lengths, are set during system design

but the filter coefficients are determined for the actual signals that shall be transmitted.

The filter coefficients $h_a(k)$ are derived according to the *Minimum Mean Square Error* (MMSE) criterion (cf. Section 2.4.2) for the prediction error signals from Equation 4.4

$$\mathrm{E}\left\{z_a(k)^2\right\} \rightarrow \min . \tag{4.17}$$

This expected value can be expanded (with $\varphi_{yz}(\lambda) = \mathrm{E}\left\{y(k) \cdot z\left(k-\lambda\right)\right\}$) to:

$$\mathrm{E}\left\{z_a(k)^2\right\} = \varphi_{y_c y_c}(0) - 2\sum_{\lambda=0}^{L-1} h_a(\lambda)\varphi_{y_c z_\mathrm{A}}(\lambda) + \sum_{\lambda=0}^{L-1}\sum_{i=0}^{L-1} h_a(\lambda)h_a(i)\varphi_{z_\mathrm{A} z_\mathrm{A}}(i-\lambda) \tag{4.18}$$

The differentiation of the expected value of the squared prediction error with respect to the filter coefficient $h_a(k)$ and setting this derivative equal to zero gives

$$\varphi_{y_c z_\mathrm{A}}(k) = \sum_{i=0}^{L-1} h_a(i)\varphi_{z_\mathrm{A} z_\mathrm{A}}(k-i) \tag{4.19}$$

This differentiation and the calculation for all filter coefficients leads after rearranging all resulting equations to a set of equations that is very similar to the regular normal equations known from single channel linear prediction [VM06]:

$$\mathbf{Z}_{\mathrm{AA}} \cdot \mathbf{h}_a = \mathbf{Y}_{c\mathrm{A}} \tag{4.20}$$

The quadratic matrix $\mathbf{Z}_{\mathrm{AA}}$ contains the autocorrelation values $\varphi_{z_\mathrm{A} z_\mathrm{A}}(\lambda)$ of the main channel in symmetric Toeplitz structure:

$$\mathbf{Z}_{\mathrm{AA}} = \begin{pmatrix} \varphi_{z_\mathrm{A} z_\mathrm{A}}(0) & \cdots & \varphi_{z_\mathrm{A} z_\mathrm{A}}(L-1) \\ \varphi_{z_\mathrm{A} z_\mathrm{A}}(1) & \cdots & \varphi_{z_\mathrm{A} z_\mathrm{A}}(L-2) \\ \vdots & \ddots & \vdots \\ \varphi_{z_\mathrm{A} z_\mathrm{A}}(L-1) & \cdots & \varphi_{z_\mathrm{A} z_\mathrm{A}}(0) \end{pmatrix} \tag{4.21}$$

The column vector $\mathbf{h}_a$ is composed of the filter coefficients $h_a(k)$:

$$\mathbf{h}_a = \begin{pmatrix} h_a(0)\ h_a(1)\ \ldots\ h_a(L-1) \end{pmatrix}^{\mathrm{T}} \tag{4.22}$$

The difference to the single channel case lies in the column vector $\mathbf{Y}_{c\mathrm{A}}$. This vector contains the cross correlation values $\varphi_{y_c z_\mathrm{A}}(\lambda)$ between the respective input channel c and the main channel:

$$\mathbf{Y}_{c\mathrm{A}} = \Big(\varphi_{y_c z_\mathrm{A}}(0)\, \varphi_{y_c z_\mathrm{A}}(-1)\, \ldots\, \varphi_{y_c z_\mathrm{A}}(1-L) \Big)^{\mathrm{T}} \tag{4.23}$$

Different approaches for solving this set of linear equations exist. Since $\mathbf{Z}_{\mathrm{AA}}$ has Toeplitz structure, the Levinson Durbin recursion is very attractive due to its low computational complexity.

These matrices and thereby also the filter coefficients can be calculated for every sample but since many signals, especially speech signals, can at least be assumed to be short term stationary, it is only slightly detrimental to calculate the filter coefficients for every block, e.g., only every 20 ms.

4.3 Impact of Quantization and Strategies for Noise Shaping

The aforementioned perfect transmission of all signals to the receiving end is not always achievable in real world systems such as telephony or broadcasting. The main constraint in this regard is the available data rate which is usually significantly smaller than the necessary rate for a *lossless* transmission. Hence, appropriate quantizers have to be used, an overview of the history of quantization can be found in [GN98], more recent developments include work on vector quantization in different domains in [Kru10, RKV12].

Irrespective of the exact type of quantizer that is utilized, it can always be modelled as noise that is added into the system. For some cases, this quantization noise $q(k)$ can be assumed to be spectrally white and as resulting from an independent process from the signal that is quantized.

Depending on the processing chain of a predictive coding system, the effective quantization noise at the output is possibly not spectrally white any more. The analysis of the different structures is most convenient in the z-transform domain. The signals and filters in this domain are denoted by capital letters in the following. The quantizer error signals $q_\mathrm{A}(k)$ and $q_2(k)$ are random processes for which the z-transform does not exist in general. However, since the processing is done in a framewise manner, the signals are time-limited which

allows to calculate the z-transform and thus describe the relations between the different signals in the z-transform domain.

In single channel linear prediction, the spectral shape of the quantization noise can be white in the case of *closed loop* prediction which allows to make use of a *Signal-to-Noise Ratio* (SNR) gain or it can have the identical shape as the quantized signal in the case of *open loop* prediction which is favorable due to the fact that masking effects make the quantization noise less noticeable. These different spectral shapes are achieved by using different encoding structures as depicted in Figure 4.6 (cf. Section 8.3.3 in [VM06]). There is also the possibility to seamlessly fade between the two extremes that are depicted. To do this, the filter $H(z)$ in the structure around the quantizer in the closed loop setup is replaced by a modified filter $H(\frac{z}{\gamma})$ with the so-called noise shaping factor γ which allows to find a suitable trade-off between the SNR gain of the closed loop setup and the psychoacoustically advantageous open loop setup. Setting this factor to $\gamma = 1$ results in the performance of the closed loop structure while $\gamma = 0$ leads to the open loop structure.

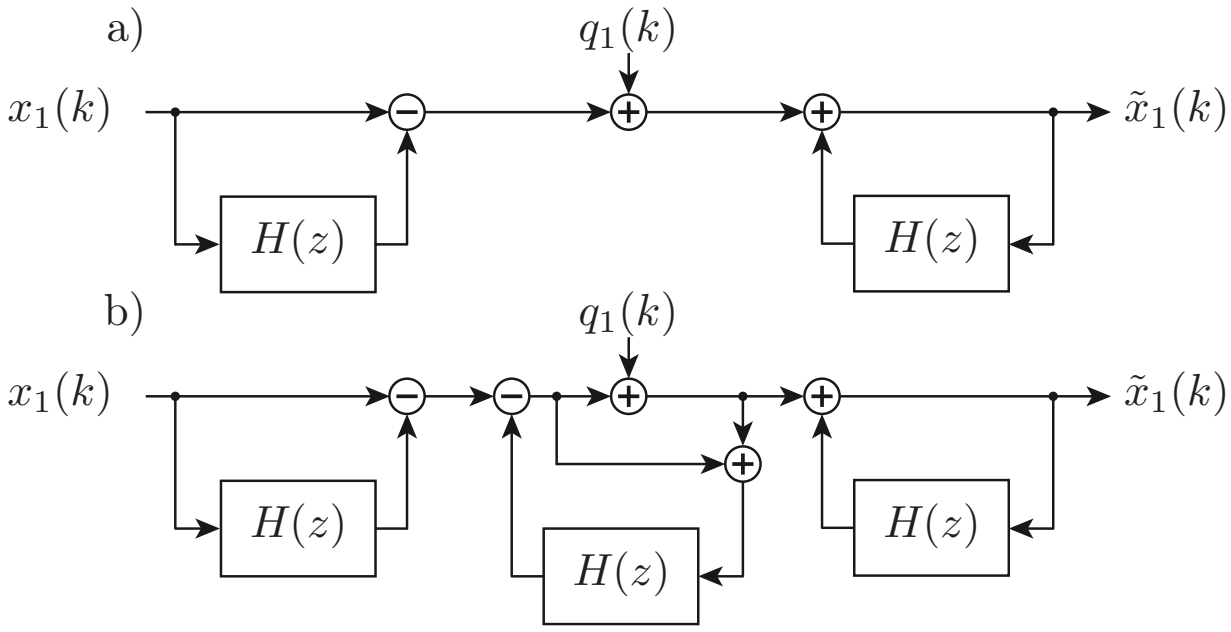


Figure 4.6: Block diagrams of open loop and closed loop single channel linear prediction

The quantity that can be used to measure the performance of the system is the effective quantization error at the different outputs. All systems that will be analyzed in this section are basically capable of perfectly reconstructing the input signals if no quantization errors are present. Hence the effective quantization error can be determined for every channel m by calculating

$$E_m(z) = \tilde{X}_m(z) - X_m(z) \cdot z^{-\tau} \,. \tag{4.24}$$

The analysis in the following is carried out for the two channel case, all the results readily extend to the multi channel case. All additional channels can be treated in analogy to the second channel that is used here. As described in Section 4.1, the only channel that is reconstructed differently is the first and all the other channels are reconstructed directly by inverting the inter channel prediction.

The naïve way to incorporate quantization into the multi channel prediction system would be to carry out all prediction processes on the transmitting side first and then quantize the signals before they are sent to the receiving side. This basic system for the transmission of stereo signals is depicted in Figure 4.7. The separation between the encoder and the decoder is represented by a vertical dashed line.

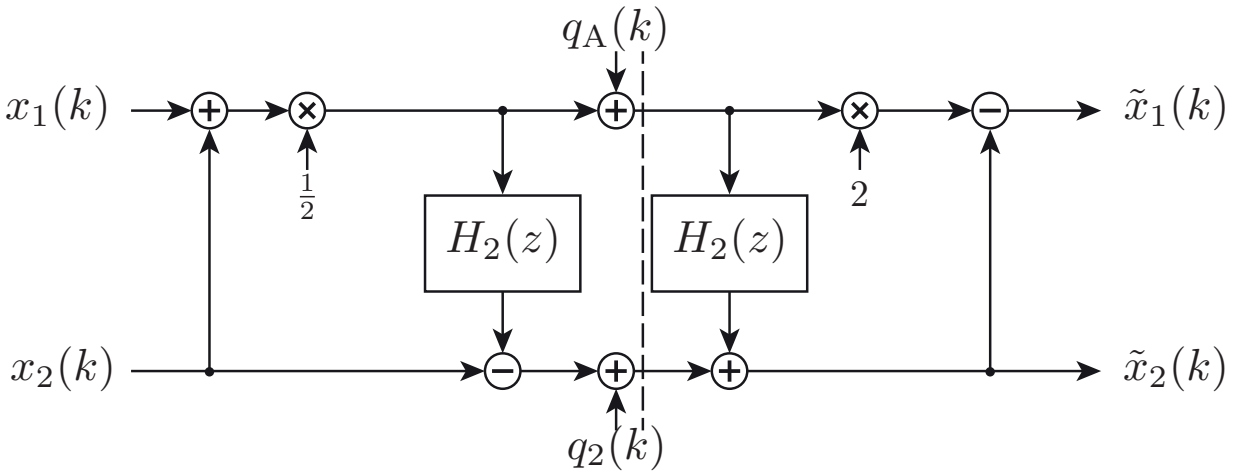


Figure 4.7: Basic System for the Transmission of Stereo Signals

The effective quantization errors in the output signals as defined in Equation 4.24 for the basic system are

$$E_1(z) = (2 - H_2(z)) \cdot Q_A(z) - Q_2(z) \tag{4.25}$$

and

$$E_2(z) = H_2(z) \cdot Q_A(z) + Q_2(z)\,. \tag{4.26}$$

The quantization error $Q_A(z)$ from the main channel is present in both outputs in different filtered versions which are, in general, not related to the spectrum of the input signals while the quantization error $Q_2(z)$ from the prediction error channel is also present in both outputs in an unfiltered form. The unwanted and not perceptually motivated noise shaping of $Q_A(z)$ can be removed by

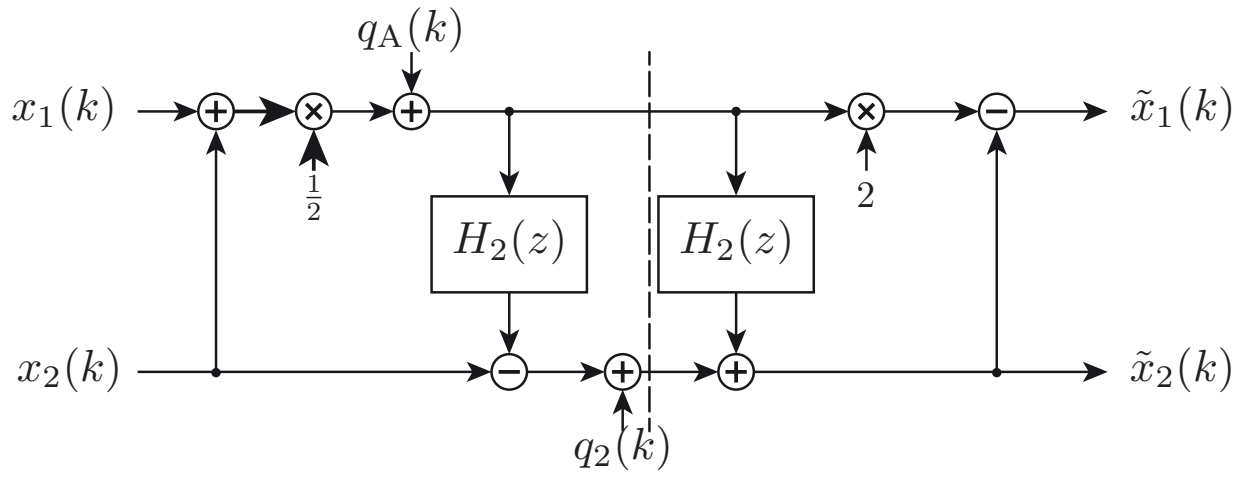


Figure 4.8: Basic System for the Transmission of Stereo Signals with moved main channel quantizer

changing the position of the quantizer in the system. The resulting, still quite basic, system is shown in Figure 4.8.

The effective quantization errors in this slightly modified system are

$$E_1(z) = 2 \cdot Q_A(z) - Q_2(z) \tag{4.27}$$

and

$$E_2(z) = Q_2(z)\,. \tag{4.28}$$

There is no spectral shaping at all, the spectra of the effective quantization errors $E_1(z)$ and $E_2(z)$ are simply unfiltered linear combinations of the spectra of the quantization noise signals $Q_A(z)$ and $Q_2(z)$. Additionally, it can be observed that the energies of the effective quantization errors in both channels are different as long as there is any quantization in the main channel A, i.e., $Q_A(z) \neq 0$. It is shown later how this asymmetry affects the performance of the entire system.

In a multi channel system, the known noise shaping schemes that are usable in single channel linear prediction are not directly applicable. It is shown in the following that the open loop setup can be combined with one variant of the multi channel prediction scheme. The closed loop case however is not as straightforward. This is obvious when looking at Figure 4.6 and the impact of the closed loop setup. The effect of whitening the effective quantization noise in the output signals results from the interaction between the open loop structure and the additional closed loop signal processing elements around the quantizer. Simply adding the identical elements to the multi channel setup does not lead to the same results since the quantizers are applied to the main channel and the prediction error and not to the input resp. output signals.

The idea behind the noise shaping modules in the single channel case is to feed the quantization error back in an appropriate manner. The same principle can also be utilized in the multi channel case. Different structures are possible, two variants are presented and analyzed here:

- A basic setup that feeds the quantization noise from the main channel to the second channel and thereby leads to spectrally white quantization errors $e_1(k) \dots e_{\mathrm{M}}(k)$ in all output channels while ensuring identical effective quantization noise energy in all channels

- A combination of the simple setup with common open loop single channel linear predictors that leads to quantization errors $e_1(k) \dots e_{\mathrm{M}}(k)$ in all output channels that are shaped like the signals in those channels

Spectrally white quantization errors with identical energy can be obtained by the structure in Figure 4.9.

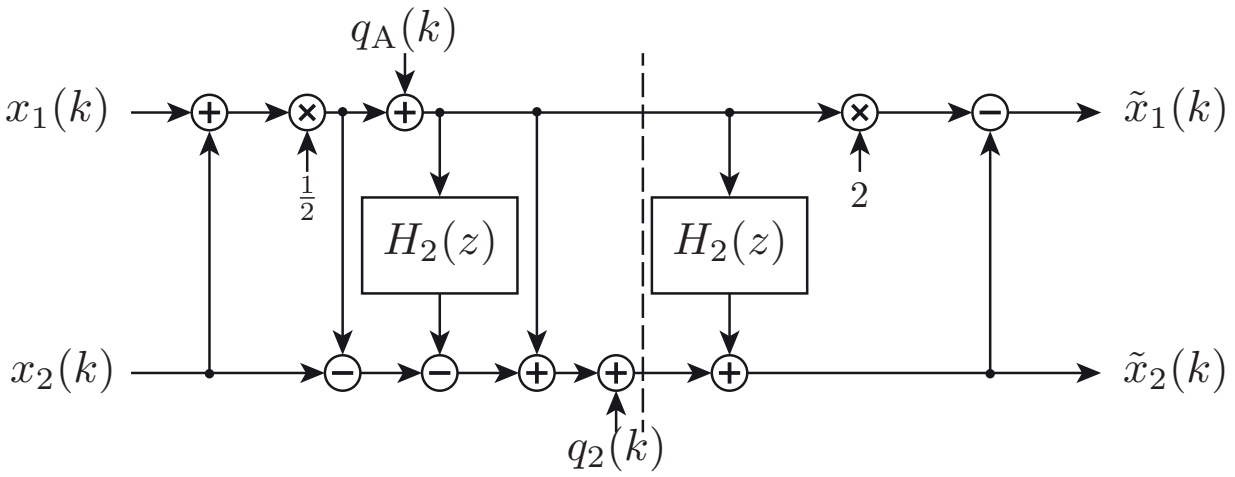


Figure 4.9: Basic Quantization Noise Feed Forward System for the Transmission of Stereo Signals

One additional summation and one additional subtraction change the effective quantization errors to

$$E_1(z) = Q_{\mathrm{A}}(z) - Q_2(z) \tag{4.29}$$

and

$$E_2(z) = Q_{\mathrm{A}}(z) + Q_2(z)\,. \tag{4.30}$$

That the two effective quantization errors have identical energy can easily be seen when calculating the difference of the two energies:

$$\begin{aligned}\mathrm{E}\left\{e_1^2(k)\right\} - \mathrm{E}\left\{e_2^2(k)\right\} &= \mathrm{E}\left\{\left(q_\mathrm{A}(k) - q_2(k)\right)^2\right\} - \mathrm{E}\left\{\left(q_\mathrm{A}(k) + q_2(k)\right)^2\right\} \\ &= 2 \cdot \Big(- \mathrm{E}\left\{q_\mathrm{A}(k) \cdot q_2(k)\right\} - \mathrm{E}\left\{q_\mathrm{A}(k) \cdot q_2(k)\right\} \Big) \\ &= -4 \cdot \mathrm{E}\left\{q_\mathrm{A}(k) \cdot q_2(k)\right\} \end{aligned} \tag{4.31}$$

Since the two quantization noises $q_\mathrm{A}(k)$ and $q_2(k)$ are uncorrelated, it follows that

$$\mathrm{E}\left\{e_1^2(k)\right\} - \mathrm{E}\left\{e_2^2(k)\right\} = 0\,. \tag{4.32}$$

The combination of the simple setup from Figure 4.9 with open loop single channel linear predictors for both input channels leads to the structure in Figure 4.10. Not surprisingly, the two single channel linear predictors shape the effective quantization error in the same way that they do in single channel linear prediction.

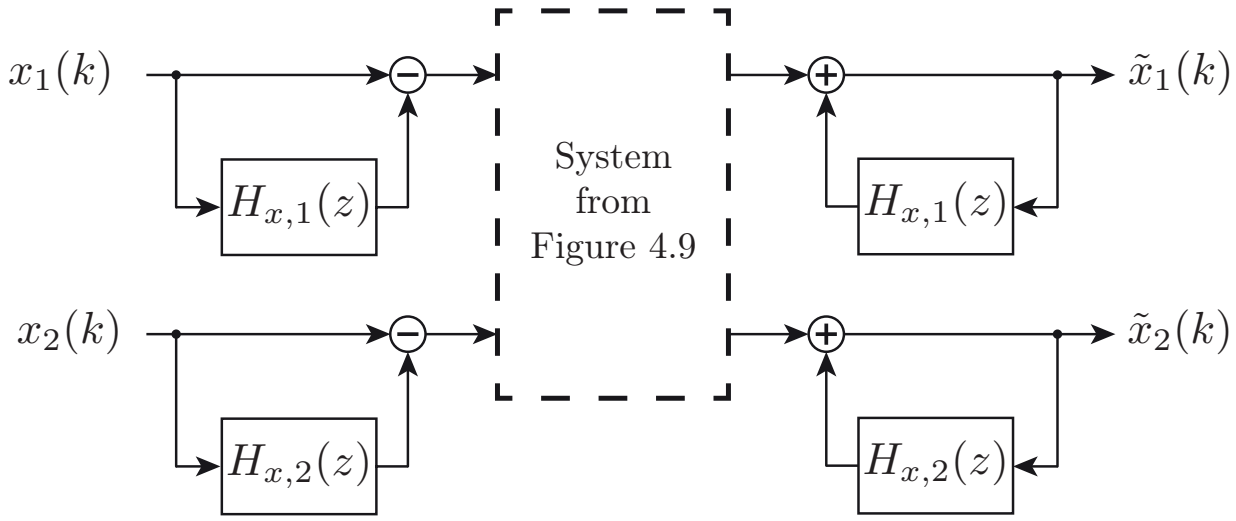


Figure 4.10: Positioning the open loop predictors around the equal quantization noise energy setup

The resulting spectral characteristics of the quantization errors at the outputs for this setup are then

$$E_1(z) = \frac{1}{1 - H_{x,1}(z)} \cdot \left(Q_\mathrm{A}(z) - Q_2(z)\right) \tag{4.33}$$

and

$$E_2(z) = \frac{1}{1 - H_{x,2}(z)} \cdot \left(Q_\mathrm{A}(z) + Q_2(z)\right)\,. \tag{4.34}$$

4.4 Experimental Evaluation

Different aspects of the presented predictive approach are evaluated in this Section. The possible designs that can be deduced from the techniques presented so far differ with respect to the signal processing structure. In addition, the specific elements have to be parameterized as well. The main focus in this evaluation is on the system dimensions, e.g., the lengths of the prediction filters and the word lengths of the utilized quantizers.

All evaluations of the performance of the presented techniques are carried out using the 3GPP stereo audio dataset [3GP07] consisting of approximately ten minutes of clear and noisy speech from various talkers in different languages as well as music signals. All signals are sampled at a sampling frequency f_s of 48 kHz. The number of channels is two for the entire dataset, hence only one set of filter coefficients and one prediction error signal has to be transmitted per frame of 20 ms besides the main channel.

4.4.1 Basic Stereo Transmission System

In a first evaluation, the stereo coding system from Figure 4.11 is considered. The filter coefficients $\mathbf{H}_2$ are assumed to be transmitted transparently (or with negligible errors, i.e., $\tilde{\mathbf{H}}_2 = \mathbf{H}_2$) while the main channel $z_\mathrm{A}(k)$ and the prediction error $z_2(k)$ are subject to quantization with the quantizers $\mathcal{Q}_\mathrm{A}$ and $\mathcal{Q}_2$, respectively.

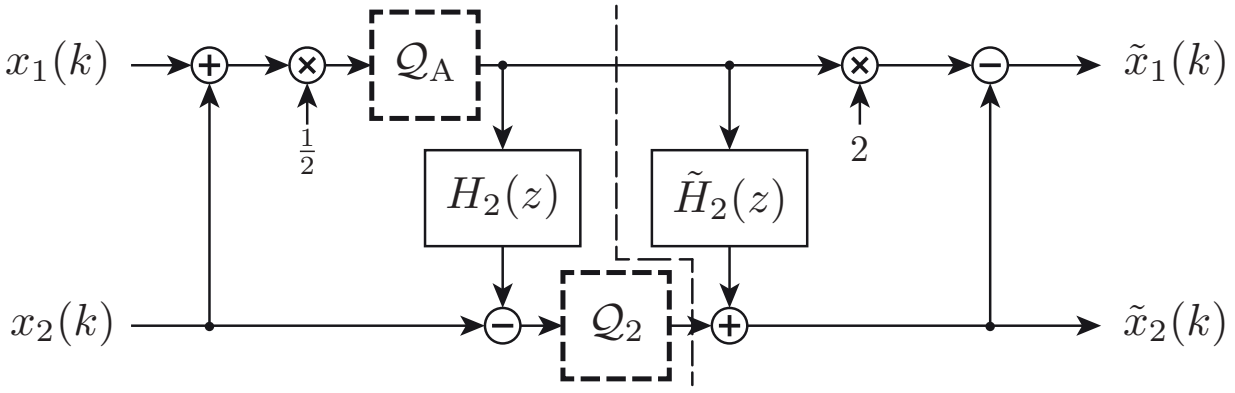


Figure 4.11: Stereo encoding and decoding system

The performance of this coding system is quantified by three measures. The results are utilized to gain insight into the choice of system parameters:

- Average prediction gain between the input signals and the prediction errors (N represents the total number of samples in the signals):

$$\overline{G_p} = \frac{1}{2} \cdot \sum_{i=1}^{2} 10 \cdot \log_{10} \left(\frac{\sum_{k=1}^{N} x_i(k)^2}{\sum_{k=1}^{N} z_i(k)^2} \right) \tag{4.35}$$

  This definition of the prediction gain differs slightly from the usually employed formula as the prediction is carried out between two signals instead of within one signal. For a practical implementation, only one prediction has to be carried out due to the symmetries described in Equations 4.7 and 4.8. However, the prediction is done for both channels here (as depicted in Figure 4.2) to increase the amount of data available for the evaluation.

- *Perceptual Evaluation of Audio Quality* (PEAQ) [ITU01a] values of the entire transmission chain. The *Objective Difference Grade* (ODG) value according to PEAQ allows to quantify the degradation of $\tilde{x}_1(k)$ and $\tilde{x}_2(k)$ with respect to $x_1(k)$ and $x_2(k)$, respectively. The PEAQ scale ranges from 0 (i.e., degradation is imperceptible) to -4 (i.e., degradation is very annoying) and a value of -2 or better is very acceptable for most use cases.

- *Advanced Objective Difference Grade* (AODG) values of the entire transmission chain. AODG is a novel instrumental quality measure that will be presented in Chapter 5, its results are already used here. It is closely related to PEAQ but is clearly superior at considering spatial properties of the signals. The scale for this measure is identical to the PEAQ scale.

Both quantizers $\mathcal{Q}_\mathrm{A}$ and $\mathcal{Q}_2$ are logarithmic scalar quantizers and the μ-law characteristic [VM06] with $\mu = 255$ is used with varying word length w from 1 to 12 bit. This very simple type of quantizer is chosen to lay the focus of the evaluation on the performance of the predictive coding step, more sophisticated quantizers that are specifically trained or adapted to the signal can easily be combined with the proposed predictive scheme to further increase performance.

The logarithmic quantizers use a uniform *mid-tread quantizer* with a symmethric characteristic (i.e., utilizing $2^w - 1$ quantization levels) as their core quantizer. This type of quantizer was chosen since it was found from informal listening tests that it is favorable for short word lengths if a value of zero for the prediction error signal can be correctly represented. Especially for a word length of 1 bit, this choice of quantizer means that all values of the prediction error are quantized to zero and only the main channel and the filter coefficients are used for the reconstruction of the input signals. The quantizer is designed

to cover the entire possible range of values of the input data (i.e., −1 to 1 for this data set).

The system design parameter of the prediction step that is evaluated first is the length L of the prediction filter which is varied from 1 to 50 taps (for a block length of 20 ms) to evaluate the impact of this variable on the overall performance of the coding system. The relation between the length L of the prediction filters, the achievable prediction gain and the transmission quality is depicted in Figure 4.12.

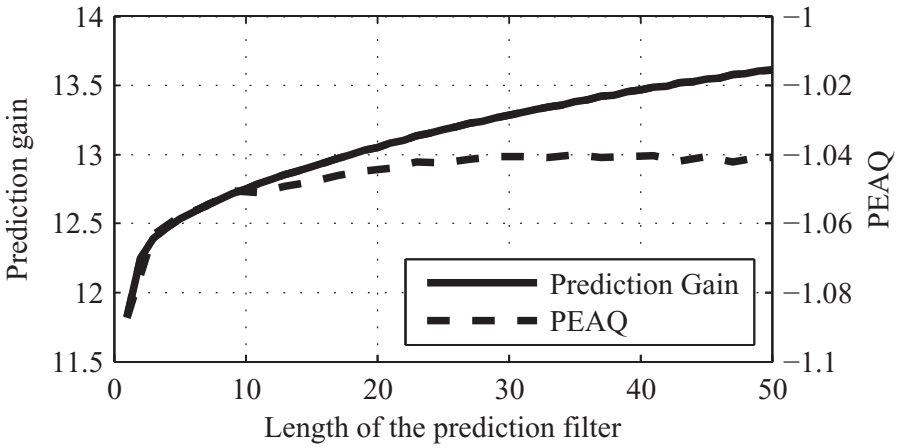


Figure 4.12: Prediction Gain and PEAQ values for different filter lengths.

It can be seen that even fairly short filter lengths offer significant prediction gains of more than 11 dB. This gain increases with increasing filter length but the increase in gain starts to get smaller fairly quickly, even a filter length of 50 taps increases the prediction gain only by roughly another 2 dB.

The average PEAQ value increases as well for an increasing length of the prediction filter. The simulations for this graph were carried out for a perfect transmission of the main channel and the results are averaged over varying word lengths for the quantizer $\mathcal{Q}_2$ between 4 and 12 bit to get one overall view on the performance of the system. A more detailed analysis of follows later on. Taking both measures into account, a filter length of between 5 and 15 taps appears to be reasonable to ensure a good performance of the coding system. This equates to an additional algorithmic delay $\frac{\tau}{f_s}$ between 0.1 and 0.3 ms.

The achievable prediction gain for $L = 11$ is depicted in Figure 4.13 over the word length of the quantizer for the main channel. The dashed line represents the prediction gain for the unquantized case as a reference.

It can be seen that the performance of the prediction step strongly depends on the chosen quantizer $\mathcal{Q}_\mathrm{A}$ for the main channel. Word lengths of less than 5 bit

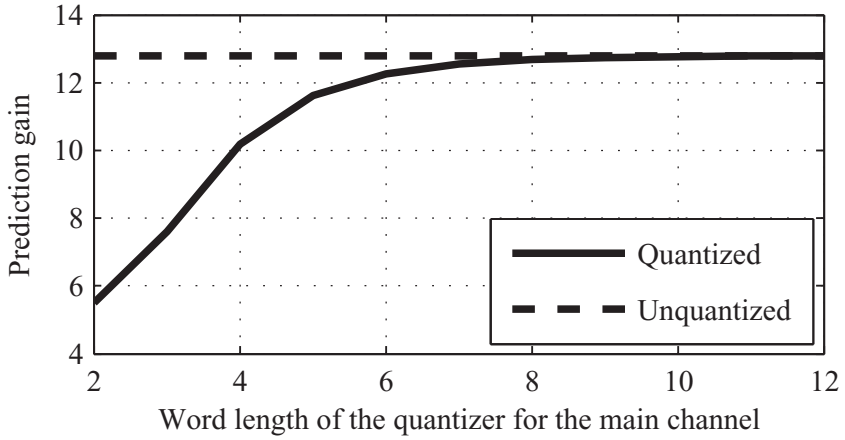


Figure 4.13: Average prediction gain for different word lengths of the quantizer for the main channel.

lead to a significant performance decrease compared to the prediction gain for the unquantized case.

The achievable quality for the transmission system of Figure 4.11 is illustrated in Figure 4.14. The abscissa therein is the word length w_2 of the quantizer for the prediction error while the set of curves consists of the different word lengths w_A of the quantizer for the main signal (2, 3, 5, 7 and 10 bit from bottom to top). The impact of the word length of $\mathcal{Q}_2$ is obviously bigger than the impact of the word length of $\mathcal{Q}_\mathrm{A}$ which can be explained by the fact that any quantization error within $z_2(k)$ will be present in both $\tilde{x}_1(k)$ and $\tilde{x}_2(k)$ without any filtering while the quantization error from the main channel is only present in $\tilde{x}_1(k)$ (cf. Equation 4.27 and Equation 4.28).

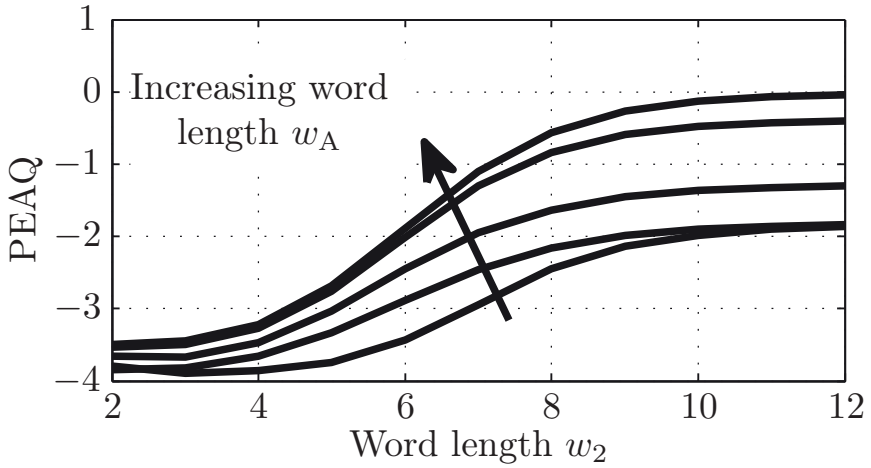


Figure 4.14: Average PEAQ values for different word lengths of the quantizers for the main channel and the prediction error. The set of curves depicts word lengths w_A of 2, 3, 5, 7, and 10 bit from bottom to top.

As a reference, the perceptual quality of a symmetric, independent quantization of $x_1(k)$ and $x_2(k)$ as depicted in Figure 4.15 is analyzed as well.

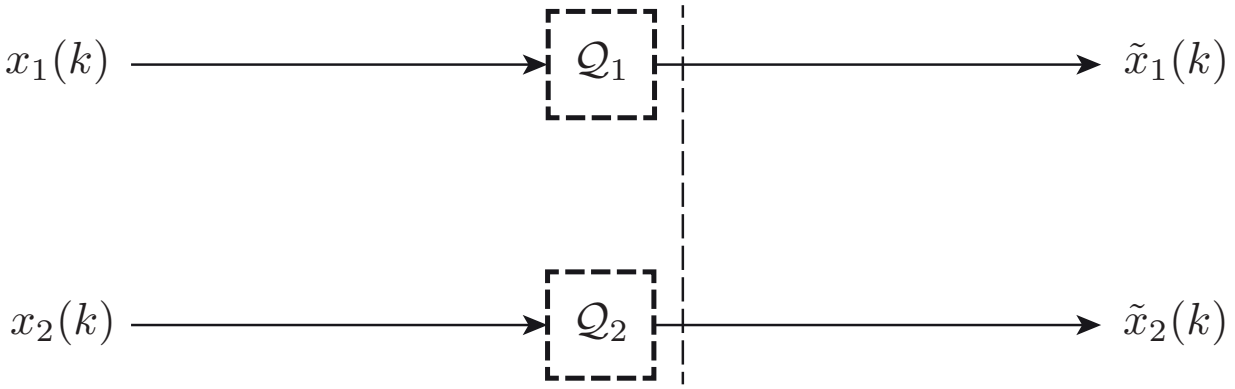


Figure 4.15: Symmetric stereo transmission system

This system uses one logarithmic quantizer for each channel that are both set to identical word lengths. It can be seen in Figure 4.16 that a longer word length leads to a higher quality and that a word length of 7 bit for each quantizer leads to a PEAQ value of approximately -2 (which is very acceptable for many use cases).

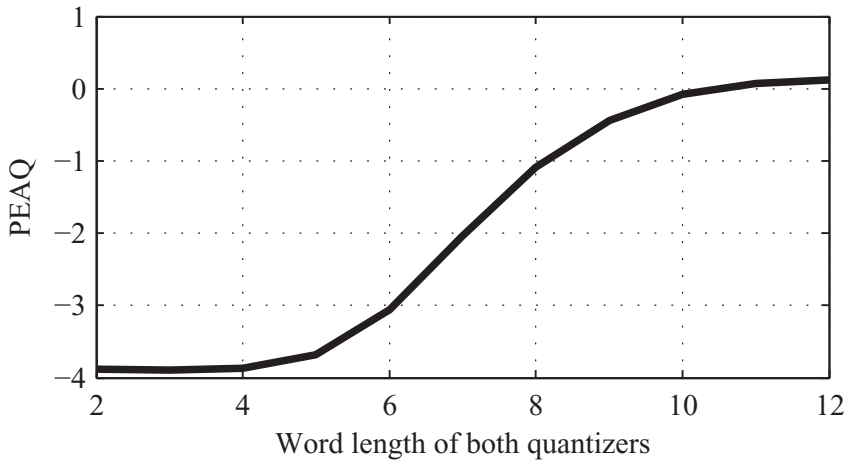


Figure 4.16: Average PEAQ values for different word lengths of the quantizers when using a symmetric transmission according to Figure 4.15.

A comparable quality for the proposed predictive coding scheme can already be reached for a combination of $w_\mathrm{A} = 2$ bit and $w_2 = 10$ bit. Including a very precise quantization of the filter coefficients (i.e., 16 bit per coefficient), the overall data rate for the proposed structure amounts to just 88 percent of the rate that is necessary for the independent transmission.

The same findings can also be made when using a single channel codec, e.g., ITU G.726 [ITU90], instead of the logarithmic quantizer. However, the evaluation

of the performance of the prediction step is possible in more detail without having a sophisticated core transmission system that masks the behaviour of the proposed audio coding structure.

4.4.2 Predictive Stereo Transmission with Equal Quantization Noise Energy

While the average PEAQ values found in the previous section for the structure in Figure 4.11 suggest a smooth and continuous quality increase when increasing the data rate, a separate analysis of the left and right channel of the audio signals shows a different picture. Due to the different energies of the effective quantization errors (cf. Equations 4.27 and 4.28), the quality of the transmission is significantly worse for $x_1(k)$. The results of separate evaluations for both channels are depicted in Figure 4.17. The results are averaged for quantizer wordlengths from 4 to 10 bit for w_A and plotted over the wordlength of w_2.

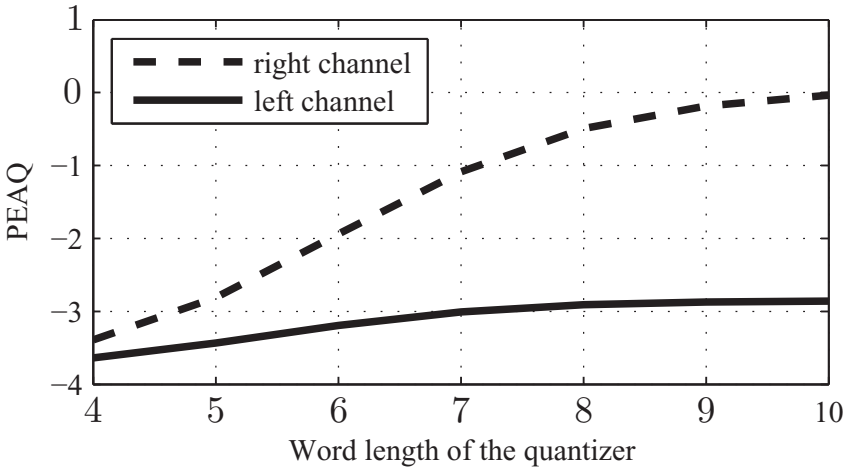


Figure 4.17: Example for the different PEAQ values for $x_1(k)$ (left channel) and $x_2(k)$ (right channel) with the setup without noise feed forward as depicted in Figure 4.7.

When the system according to Figure 4.8 that ensures equal noise energy in both channels is used, the performance of the system is much more symmetrical as illustrated by Figure 4.18.

The overall performance of the system is again illustrated by the average PEAQ and AODG results which are plotted over the word length w_2 of the quantizer for the second channel and every curve represents one word length for the main channel quantizer (from bottom to top: $w_\mathrm{A} = 2$, 3, 5, 7, and 10 bit). The PEAQ results are depicted in Figure 4.19. It can be seen that the overall

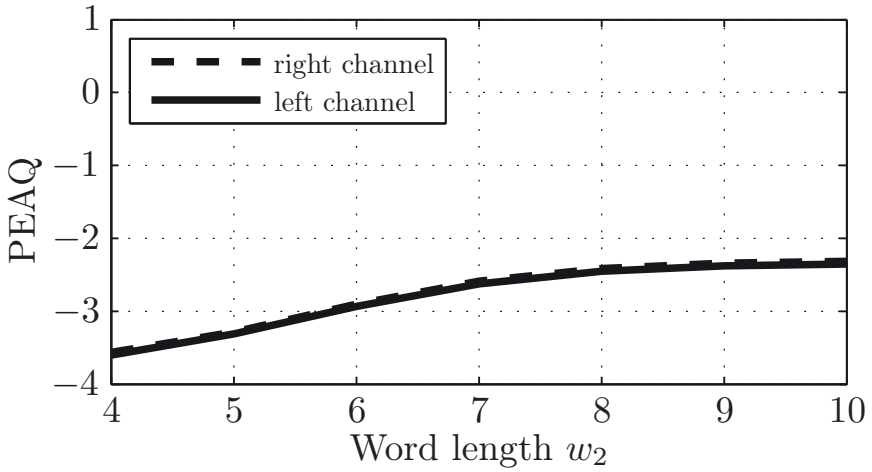


Figure 4.18: Example for the more similar PEAQ values for $x_1(k)$ (left channel) and $x_2(k)$ (right channel) in the setup with noise feed forward as depicted in Figure 4.8.

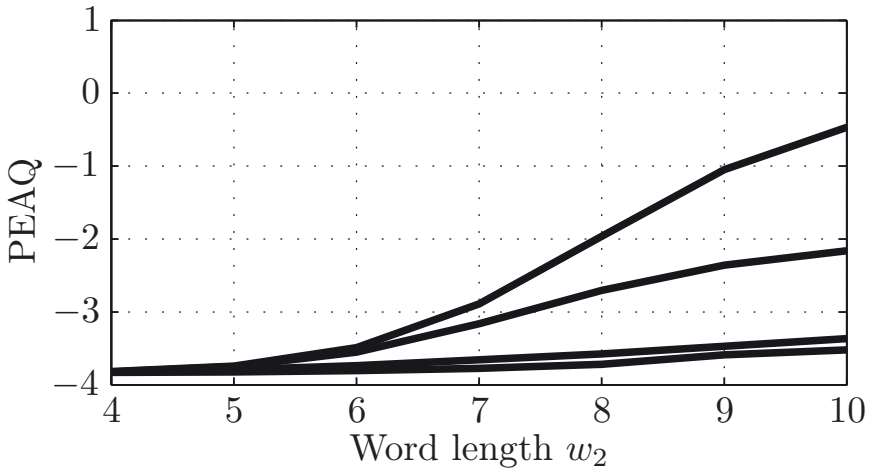


Figure 4.19: PEAQ values for the setup with noise feed forward. The set of curves depicts word lengths w_A of 2, 3, 5, 7, and 10 bit from bottom to top, the curves for 2 and 3 bit are almost identical in this resolution.

performance now strongly depends on the word length w_A of the main channel quantizer and that a good transmission quality with respect to this measure is only possible if both quantizers have a reasonable word length.

The novel quality measure AODG is introduced in the next chapter. It will be shown to be superior to PEAQ with respect to the handling of stereo signals by also integrating spatial properties of the signals. The results for this measure are visible in Figure 4.20.

The results for this measure show a saturation behaviour that depends on the word length of the main channel quantizer. A good quality is achievable for fairly low data rates already.

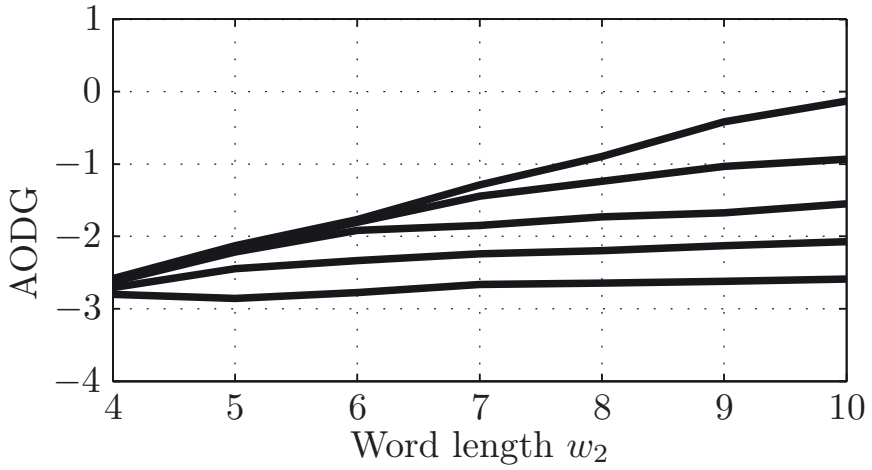


Figure 4.20: AODG values for the setup with noise feed forward. The set of curves depicts word lengths w_A of 2, 3, 5, 7, and 10 bit from bottom to top.

4.4.3 Stereo Transmission with Open Loop Noise Shaping

From a perceptual viewpoint, it would be advantageous to additionally include open loop predictors (cf. Section 2.4.1) for the input signals as well as shown in Figure 4.10. As described in Section 4.3, the system with equal quantization noise energy in both channels according to Figure 4.9 provides the basis for this. The same instrumental measures as before are used to analyze the behaviour of the system. First, the PEAQ results are shown in Figure 4.21. In comparison

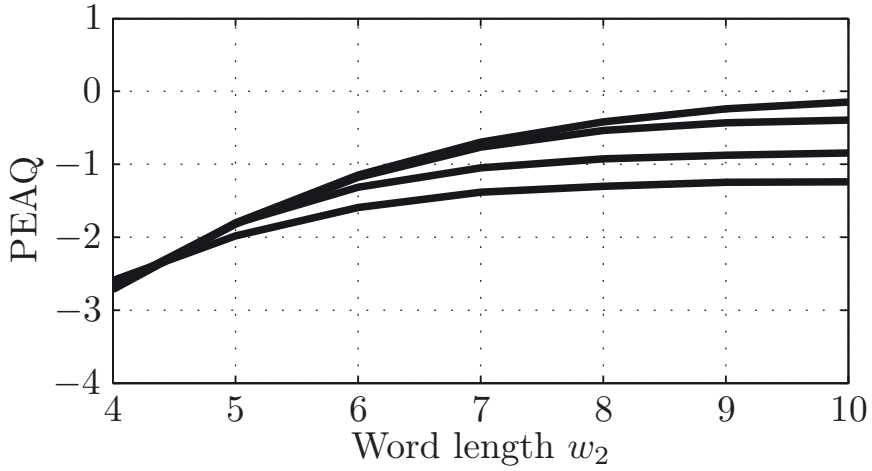


Figure 4.21: PEAQ values for the setup with noise feed forward and additional open loop noise shaping. The set of curves depicts word lengths w_A of 2, 3, 5, 7, and 10 bit from bottom to top, the curves for 2 and 3 bit are almost identical in this resolution.

to the system without the open loop predictors (cf. Figure 4.19), the strong

effect of the open loop predictors is clearly visible. Especially for lower word lengths of either the main channel quantizer or the quantizer for the residual signal, the performance is significantly better.

The AODG results are depicted in Figure 4.22. The same trend that could be

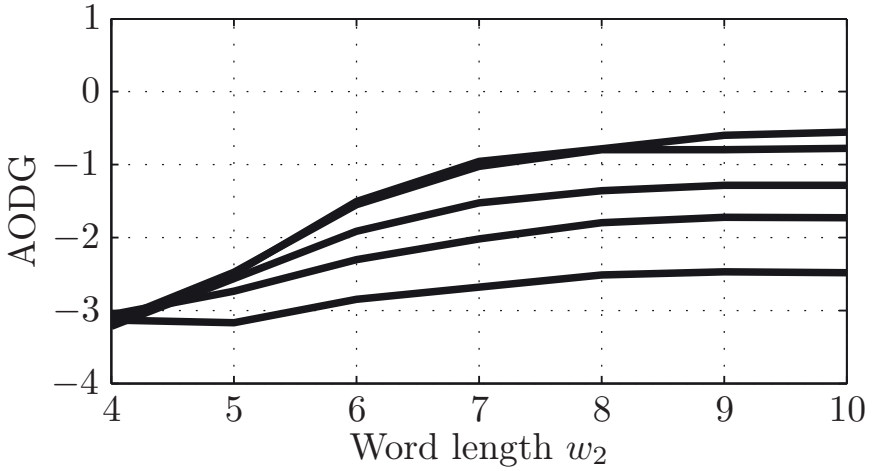


Figure 4.22: AODG values for the setup with noise feed forward and additional open loop noise shaping. The set of curves depicts word lengths w_A of 2, 3, 5, 7, and 10 bit from bottom to top.

seen when comparing the PEAQ values is again obvious: The addition of the open loop predictors for the input signals improves the performance especially for lower quantizer word lengths.

4.5 Application Example

As an application example, the combination of the proposed prediction concept with an independent core codec for the transmission of the downmix and the prediction error signals is presented in this section. The core codec in question here is the *Adaptive Multi-Rate Wideband* (AMR-WB) codec standardized by both ITU [ITU03] and 3GPP [ETS09].

The combination of the core codec with the predictive coding scheme is very straightforward: The main channel signal $z_\mathrm{A}(k)$ and the prediction error signal $z_2(k)$ are encoded by separate instances of the core codec and transmitted. Depending on the exact setup, a local decoding has to take place, e.g., if exactly the setup from Figure 4.11 shall be used in combination with a single channel core codec, the resulting system will contain one local decoder for the main channel signal.

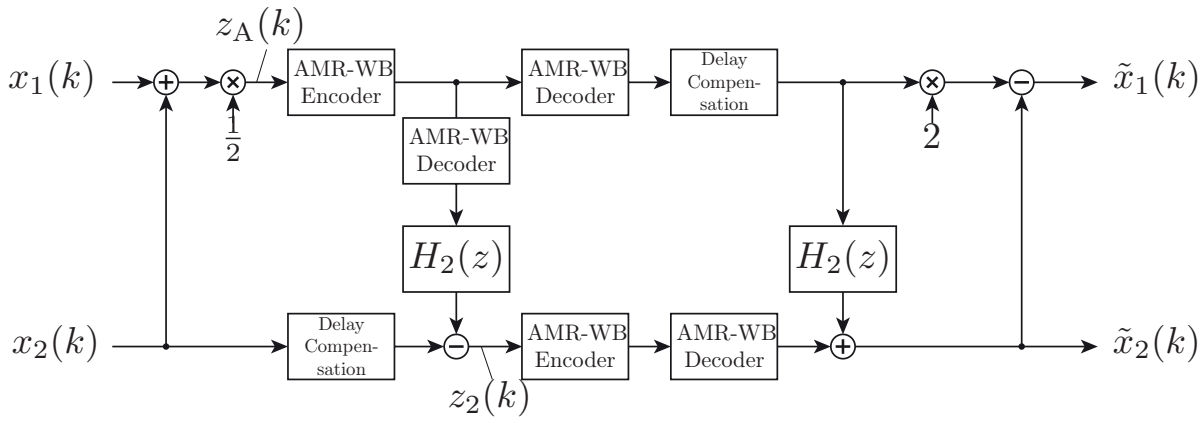


Figure 4.23: Stereo encoding and decoding system

The system can be extended in analogy to the techniques presented in Section 4.3 to achieve certain noise shaping characteristics. It has to mentioned that the noise of a codec in general and the internal noise of AMR-WB in particular is definitely not spectrally flat. Hence it is not useful to take the presented noise feed forward and noise shaping techniques (cf. Figures 4.8, 4.9 and 4.10) into consideration here. They are sensible if a completely new codec shall be designed on this multi channel predictive approach.

One aspect of this setup that has to receive a closer look is the algorithmic delay. In the basic setup of the predictive coding scheme as described in Section 4.1, the input signal $x_2(k)$ is delayed just by τ (cf. Equation 4.3) to allow for a symmetric filter impact. When the local decoding is in place as depicted in Figure 4.23, the delay of all input channels has to be increased by the algorithmic delay of the codec so that the additional delay of the multi channel extension is no longer negligible. Due to this delay reason, a system as depicted in Figure 4.24 is used for the evaluation here.

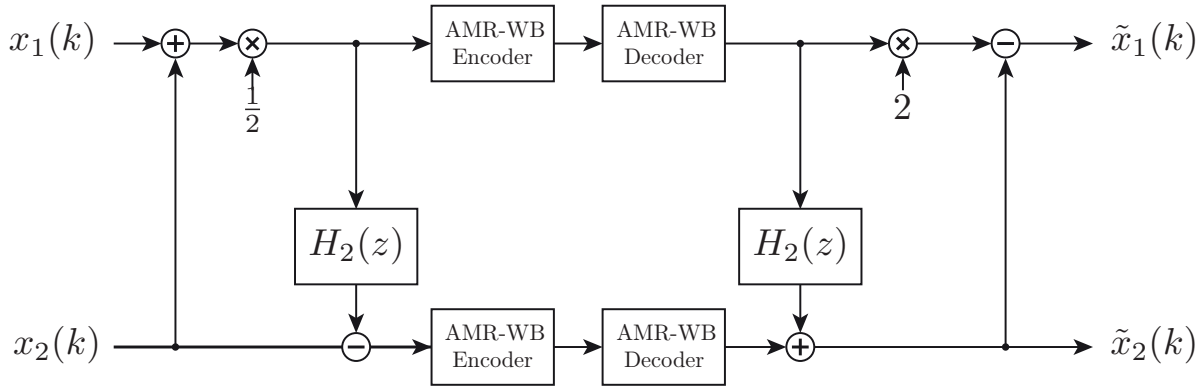


Figure 4.24: Stereo encoding and decoding system

This structure is equivalent to the basic system of the different structures that are presented in Section 4.3. The considerations about the spectral shape of the

quantization noise are not applicable here since the noise that is introduced by the AMR-WB is not white. Nevertheless, it has to be evaluated depending on the application scenario if the additional delay of the system from Figure 4.23 or the noise filtering of the system from Figure 4.24 is more acceptable.

AMR-WB has the advantageous property that it offers 9 different data rates. When combining it with the predictive coding scheme, these can be used to facilitate different data rate distributions between the main channel and the prediction error signal by using two individual instances of AMR-WB for these two channels.

In [SKV09], the overall stereo coding performance of the new approach was evaluated by varying the data rates for both the main channel and the prediction error signal. As as reference, independent transmissions of the two input channels and M/S joint-stereo coding were also considered and the same quality as for these references could be reached at a significantly lower data rate with the predictive coding scheme. Compared to the independent transmission, just 53.4 % of the data rate is necessary and even in comparison to the M/S joint-stereo coding, a decrease by 31 % is possible.

The aspect of the transmission of the filter coefficients has to be considered as well for a practical integration of the stereo extension. One attractive possibility to transmit the prediction filter coefficients is to use the data hiding scheme from, e.g., [GV07] as also used in [GSV11] for transmitting binaural wideband signals over a narrowband codec.

4.6 Conclusions

In this chapter, a concept for the transmission of multi channel signals was introduced and analyzed. The concept is based on a predictive coding system that uses a main channel signal as the basis for predicting the other channels. It was shown to add just a few samples of algorithmic delay and it is well suited for combination with common existing mono core codecs. Due to the main channel signal, the concept offers inherent backwards compatibility. In a heterogeneous communication network consisting of single channel as well as multi channel devices, this main channel can directly be provided to the communication partners with a single channel device.

The presented encoding structure achieves high prediction gains of more than 12 dB already for filter lengths $L \geq 2$. A detailed evaluation of the impact of scalar quantization on the overall system performance was carried out based

on the instrumental measures PEAQ and AODG. Both instrumental measures show that the proposed audio coding system achieves a very good transmission quality at higher data rates and that there is a graceful degradation in transmission performance through medium data rates. The evaluation also verified that the data rate efficiency for all variants of the proposed system is significantly higher than for an independent transmission of the input channels which uses the same quantizers.

Novel techniques for noise shaping in a multi channel scenario were devised. They are based on feeding the quantization noise from the main channel to the prediction error signal channels. A combination with single channel open loop predictors for the input channels was shown to achieve the highest transmission quality.

5

Instrumental Quality Measure for Multi Channel Systems

When designing speech and audio signal transmission or enhancement systems, it is important to evaluate the perceived quality. The gold standard for this are listening tests [ITU96a] which are very flexible as they can be tailored to the *System under Test* (SUT). There are tools available to conduct listening tests (e.g., [SSGV11]). During algorithm development however, continuously conducting listening tests for each minor modification is too time consuming.

Hence, the instrumental assessment of the perceived quality is a topic that has been receiving continuous interest. An overview on quality assessment in general can be found in [RBK+06, Côl1]. Basic approaches for the evaluation of multi channel signals in particular are considered in [GZR06, ZRKB05].

A novel, more advanced concept for the quality evaluation of multi channel signal processing systems is described in the following. It is based on an improved coincidence-based binaural hearing model which consists of a physiologically motivated signal processing step and a subsequent cognitive model. The hearing model is shown to be capable of correctly detecting and tracking sources even in adverse acoustic environments. Spatial parameters are derived from this model which are then combined with the output of a known algorithm for the evaluation of audio quality to get a joint measure for audio quality and spatial fidelity.

Besides the spatial parameters, the model can also blindly determine the number of active sources based on temporal and spectral information which makes it a logical enhancement of source separation algorithms that rely on this knowledge.

Without the additional parts for quality evaluation, the hearing model alone is of interest in all areas that need to consider the binaural capabilities of the human hearing system. Of the various parameters that can be derived from the hearing model, the five most important spatial parameters for the perceived quality are determined based on listening tests and subsequently utilized for an quality measure. This measure is derived with a methodology that is very similar to the basis for the *Perceptual Evaluation of Audio Quality* (PEAQ) measure.

The instrumental quality measure which is described and evaluated in this chapter as an extension to PEAQ was first introduced in [SBV13] and utilizes the binaural auditory and perceptual model from [SBV12]. The quality measure consists of the calculation of binaural *Model Output Variables* (MOVs) which are fed to a *Neural Network* (NN) together with the result of PEAQ to get the overall result.

Instrumental evaluation of the perceived audio signal quality is an important tool for the development of audio signal enhancement and transmission systems. There are various single channel measures which can be used for different application scenarios. Binaural signals have not received much attention so far and no sophisticated model of spatial perception is utilized in the available measures. It is shown that the inclusion of spatial information into the instrumental quality by the presented *Advanced Objective Difference Grade* (AODG) measurement leads to a strongly increased correlation between the instrumental measure and a listening test.

This chapter is structured as follows: After a short review of known instrumental measures in Section 5.1, the general concept of the binaural hearing model and the novel extensions are described in Sections 5.2 and 5.3. In Section 5.4, the clustering procedure along with the new cognitive strategy to estimate the number of sources is introduced followed by the experimental setup and a discussion of the results. The concept of the PEAQ add-on and the parameters that are derived from the binaural hearing model are described in Section 5.5. These parameters are mapped to the final quality measure, as described in Section 5.6. The capabilities of the quality measure are evaluated in Section 5.7 before concluding remarks are given in Section 5.8.

5.1 Known Instrumental Measures

To decrease the necessary effort for determining the quality of the transmission, various instrumental measures have been devised and used for a long time. The

Signal-to-Noise Ratio (SNR) between the energy of the transmitted signal and the energy of the noise that was introduced by the transmission (e.g., due to quantization) is a very rough estimate of the quality of a transmission system.

For high SNR values, this estimate correlates very well with the subjective impression. However, for medium to low SNR values, the correlation is significantly lower among other things due to the fact that the SNR includes no model of the human perception and weights all noise parts identically. Depending on the structure of the transmitted signal and the noise, effects like masking can lead to drastically different perceptions for identical SNRs.

More sophisticated systems for the evaluation of signal processing systems like PESQ [ITU01b, ITU05] and POLQA [ITU11] have been established for single channel speech enhancement and transmission systems. For generic audio signals, PEAQ [ITU01a] is a reliable measure. Therefore, it is often used in different areas of acoustic signal processing [VTMM10, SV12b]. The fundamental principle of PEAQ is the calculation of so-called *Model Output Variables* (MOVs) of a monaural hearing model, comparing these MOVs of the reference (input) signal and the degraded (output) signal of the SUT and feeding these differences into a NN that is trained based on the known results of numerous listening tests. The final output is a value on the *Objective Difference Grade* (ODG) scale which ranges from 0 (no audible degradation) to -4 (very annoying degradation).

PEAQ does offer the possibility to evaluate stereo signals as well. In this case, two monaural hearing models are used in parallel ([ITU01a]: "... in the case of stereo signals all computations are performed in the same manner and independently of one for the left and right channel."), i.e., no inter-channel cues are taken into consideration. The MOVs of the two channels are then averaged before the NN and an overall quality for the stereo signal processing system results. The effect of this averaging before the NN is very similar to just using PEAQ separately for the two channels and averaging the final ODG values.

It will be shown in this chapter that this leads to a fairly poor matching between the estimated quality and the perceived subjective spatial quality for many audio signals. Many lossy transmission systems using speech or audio codecs are not able to exactly preserve the positions of the various sources within the auditory scene or they may even discard the spatial information altogether if the available data rate is too small. Since the fidelity of this spatial information is not explicitly considered in PEAQ, this may lead to faulty quality estimates. A binaural hearing model that can be used to quantify the spatial properties of the signals is introduced first before the discussion returns to the actual quality measure and its integration into PEAQ.

5.2 Spatial Hearing Models

Spatial hearing is a research topic that has received continuous interest from different scientific areas throughout the last century with Rayleigh's duplex theory [JWS07] as an important first milestone highlighting the role of *Interaural Time Difference* (ITD) and *Interaural Level Difference* (ILD). Throughout the 20^{th} century, many research efforts concentrated on descriptive evaluations of the properties and capabilities of the human hearing system, both for monaural and binaural perception – overviews on many of the experiments that were carried out can be found in [FZ07, Bla97].

The next major step towards accurately modeling the way that humans perceive sounds in general and spatial properties in particular took place after also considering the increasing knowledge about human physiology. An overview on the way that acoustic events are processed in the human auditory system and many of the derived models can be found in [Bod95].

The different binaural perception models are mostly derived from two fundamental theories:

- The coincidence-based model of Jeffress [Jef48]
- The equalization-cancellation model of Durlach [Dur63]

Most modern models of binaural hearing are based on Jeffress's introduction of special neural units called *coincidence cells*. These record coincidences in neural firings from hair cells from both ears within one frequency band. Furthermore, the neural signals are delayed by a small amount that is fixed for a given fiber pair. In other words, the ITD of a single stimulus will be coded by means of a so-called *internal time difference* τ. This specific τ refers to the coincidence cell having the highest response activity (highest fire rate). By finding the most common internal time difference over a certain frequency range, the direction of the source can be estimated. Jeffress's model is actually a coarse representation of the structure of the auditory nervous system, which was unknown at that time and could only be detected physiologically significantly later [YC90].

The equalization-cancellation model was originally designed for modeling binaural masking level differences by first aligning the ear signals and then subtracting them from one another. The residual signal can then be interpreted to gain insight into the behaviour of the human hearing system. It was later shown that it is a feasible way to explain various other auditory phenomena as

well. The capabilities and limitations of this model are discussed in detail in [CD78].

In the following sections, the coincidence-based model and its extensions [Bla05] will be improved by incorporating both a weighting function for different subbands as well as a distribution function of the density of the hair cells. It is shown that these steps, when combined with skeleton correlation as a post-processing step, considerably improve the results for complex auditory events.

In addition to these improvements of the binaural auditory model, a new cognitive model is utilized which attempts to replicate the processes that are carried out by the human auditory system to estimate both the number of active sources as well as their direction. The combined system can then be used, e.g., as the preliminary stage of any source separation method. There are a variety of methods that perform blind source separation. However, the number of sources are usually assumed to be known. Unlike the method which is presented in [Arb07], the focus of this system is to replicate the ability of the human auditory system.

5.3 Binaural Hearing Models

Mathematically speaking, the fundamental element of all the coincidence-based models is a short-time cross-correlation of the neural signals $n_l(k)$ and $n_r(k)$ (with the discrete time index k) that are produced by the hair cells in the cochleas which are the final stage of monaural hearing models that are used for both ear signals $\tilde{x}_l(k)$ and $\tilde{x}_r(k)$. A disadvantage of the basic coincidence-based model is that in the localization, only the ITD is taken into account. Given that our brain mostly utilizes the ILD to locate a sound event for frequencies greater than 1500 Hz [JWS07], this is a major drawback.

To overcome this issue, two major extensions to the coincidence model were proposed by Lindemann in [Lin86a, Lin86b]. First, monaural detectors are included in the model to ensure that the model output is sensible for monaural or near-monaural cases (i.e., the level of the signal at one ear is negligible compared to the other one). The most important extension, however, is an inhibition mechanism that suppresses the fire rate caused by coincidence units if the fire rate of neighboring coincidence units is very high.

This so-called *contralateral inhibition* leads to sharper peaks of the correlation diagram along the internal delay axis as well as to increased sensitivity w.r.t. to the ILD. Moreover, the ambiguities about the ITD at higher frequencies are

suppressed. In the next section, the extended model is briefly reviewed. A detailed description can be found in [Lin86a, Lin86b].

5.3.1 Binaural Hearing Model According to Lindemann

The binaural hearing model can be integrated into a complete hearing model as depicted in Figure 5.1[1]. The transfer functions of the outer and middle ear are not considered explicitly as these have none or only marginal influence on spatial perception.

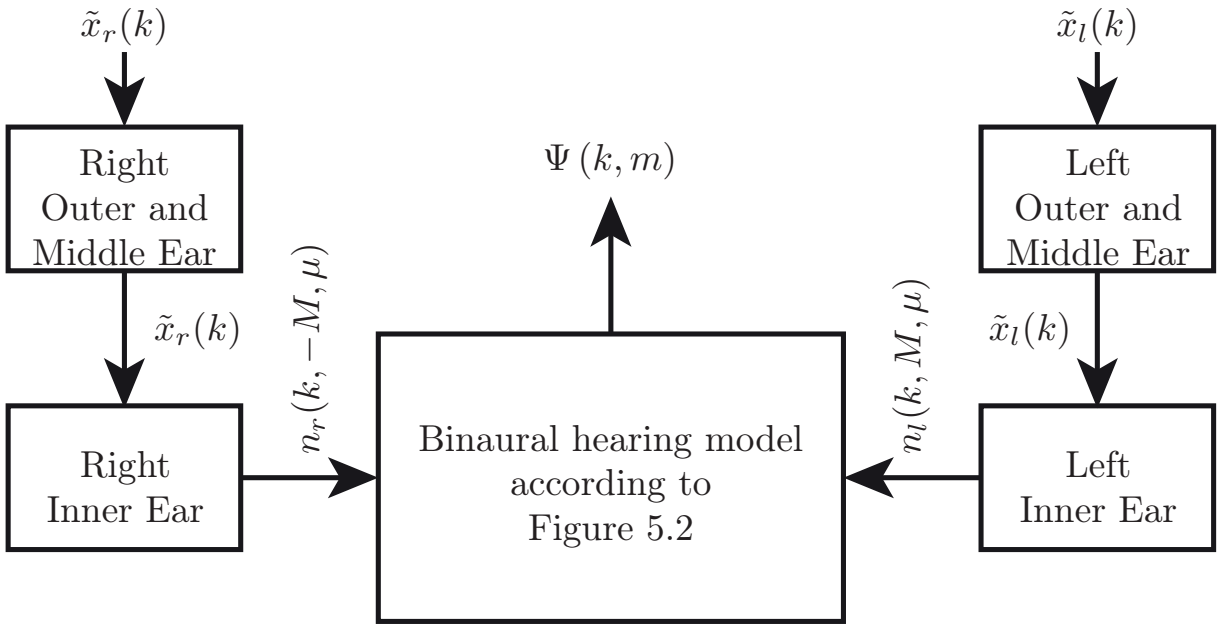


Figure 5.1: Block diagram of the complete hearing model

The signals $\tilde{x}_l(k)$ and $\tilde{x}_r(k)$ are the input of the discretized inner ear model. The inner ear was modelled in the original proposal as a 36-channel filter bank of band pass filters according to the critical bands [ZF67]. These band pass filters cover the frequency range from 20 Hz to over 16 kHz. As an improvement to the original model, the filter bank here is implemented as a Gammatone filter bank (also with 36 channels) which is known to be better adapted to the transmission function of the cochlea [HNS88].

The subband signals are then subject to a half-wave rectification, a square root function and a low pass filter with a cutoff frequency of 800 Hz to complete the transfer function of the hair cells within the cochlea. This allows to extract

[1]The reason for placing the right ear on the left side and vice versa in this figure is the correct orientation of $\Psi(k,m)$ as is explained later.

the envelope of the stimulus which is needed to correctly model the capability of the human hearing system to utilize the ITD even for the localization of complex stimuli containing only high frequencies. The subsequent stages are working on a discretized τ-axis in steps of $\Delta\tau$ according to $\tau = m \cdot \Delta\tau$.

The output of this stage, i.e., the signals $n_l(k, M)$ and $n_r(k, -M)$ with $\pm M$ as the extreme values on the discretized τ-axis, is the input to the Lindemann model which consists of a bi-directional chain of delay elements $\Delta\tau$ with additional time and amplitude variable multipliers ($\otimes$). A block diagram of the model is depicted in Figure 5.2. The leftmost and rightmost channels therein (i.e., $\Psi(k, -M)$ and $\Psi(k, M)$, respectively) are the aforementioned monaural detectors which are activated by sound events at one ear for which no corresponding event at the other ear is present.

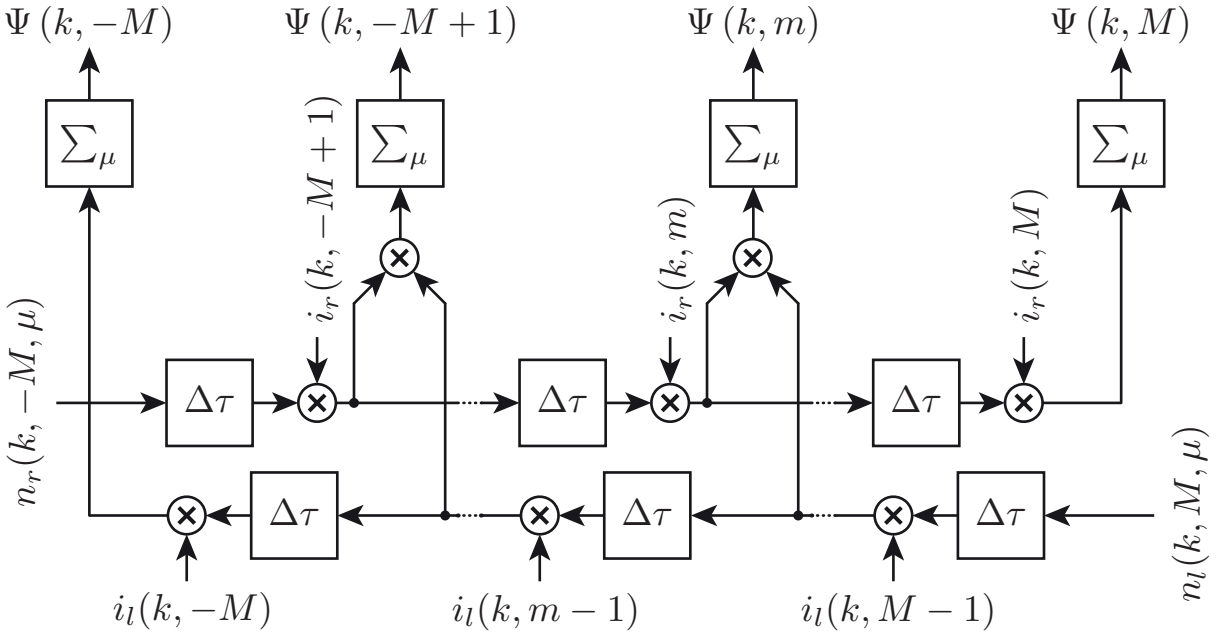


Figure 5.2: Block diagram of the binaural hearing model according to Lindemann [Lin86b]

The output of the model is the inhibited cross-correlation function $\Psi(k, m)$ which is calculated per frame as

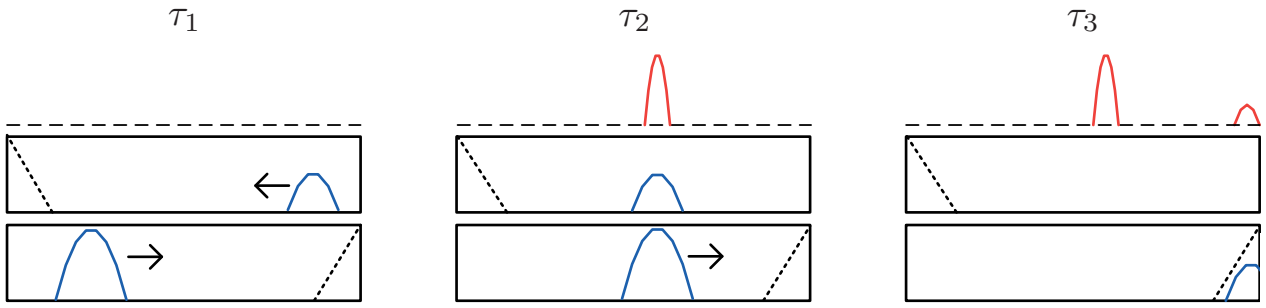
$$\Psi(k,m) = \sum_{\mu=1}^{36} \sum_{i=k-N+1}^{k} \varphi_{n_l n_r}(k,m,\mu) \cdot \mathrm{e}^{\frac{i-k}{N}} \tag{5.1}$$

with the frequency band index μ and N as the number of samples per 5 ms frame. The cross products $\varphi_{n_l n_r}(k, m, \mu)$ are obtained from the inhibited left and right neural signals $n_l(k, m, \mu)$ and $n_r(k, m, \mu)$ as follows:

$$\varphi_{n_l n_r}(k,m,\mu) = n_l(k,m,\mu) \cdot n_r(k,m,\mu) \tag{5.2}$$

The inhibition mechanism according to [Lin86b, Lin86a] has both a stationary and a dynamic component and is given by $i_l(k,m)$ and $i_r(k,m)$ in Figure 5.2 for the right and left channel, respectively. The stationary inhibition decreases the amplitude of both signals before every delay element $\Delta\tau$ along the τ-axis with respect to the contralateral signal. The dynamic inhibition is the output of a lowpass whose input signal is the cross product $\varphi_{n_l n_r}(k,m,\mu)$. This dynamic inhibition is utilized to model additional binaural properties as, e.g., the law of the first wave front.

With this inhibition mechanism, the model is also sensitive to ILD which leads to a much more realistic representation of the human capabilities. However, this can be disadvantageous for certain sound events which exhibit significantly different neural signal amplitudes at both ears and consist of only a single signal at both ears. In that case, additional peaks are generated along the τ-axis [Gai93] as illustrated by an example of a sound event from the right side in Figure 5.3.



a) neural signals have just entered the chain of delay elements

b) neural signals are at the same position in the chain of delay elements

c) remaining neural signal reaches the right monaural detector

Figure 5.3: Formation of additional peaks in $\Psi(k,m)$ (top row) due to monaural detectors for $n_l(k,m,\mu)$ (second row) and $n_l(k,m,\mu)$ (third row) – depicted for three internal delays: $\tau_1 < \tau_2 < \tau_3$

In all subplots, the bottom parts represent the bi-directional chain of delay elements and the neural signal(s) are visible as they are progressing along the chain. The monaural detectors are represented by the dashed triangles at the leftmost and rightmost end of the delay chain, respectively. Note that the neural signal n_r of the right side enters its chain from the left and is depicted in the lower part of the figure while the neural signal n_l of the left side enters from the right and is depicted in the middle. Even though this might seem like an unintuitive setup, it leads to the output of the chain (i.e., the inhibited cross-correlation function $\Psi(k,m)$ which is visible in the top part of all three

panels) being the right way around. Hence, a maximum of the inhibited cross-correlation function in the right half indicates a sound source on the right.

In the left subplot, the signals have just entered the chain of delay elements and $\Psi(k, m)$ is all zero since the respective neural signals are only entering the chain and there are no overlapping segments of the two signals yet.

In the central subplot, the signals are at the same position in the chain of delay elements and the main peak of the inhibited cross-correlation is generated. Thanks to the inhibition mechanism, the smaller neural signal is completely inhibited after this while the amplitude of the larger signal is reduced.

In the right subplot, the monaural detector on the right side is activated by the residual neural signal in the lower delay chain and an additional maximum results, falsely indicating an additional sound source all the way to the right. A solution to overcome this drawback is presented in Section 5.3.2.

5.3.2 Improved Delay and Frequency Weighting

In order to specifically suppress the additional peaks that stem from the monaural detectors, Gaik [Gai93] already presented an additional weighting of the signal and extended the model by an adaptation to the individual *Head-Related Transfer Function* (HRTF). After an extensive learning phase, the proposal from [Gai93] leads to a more natural combination of ITD and ILD with the disadvantage that this only works well for the HRTFs that were used in the training phase.

An alternative, more generic solution is proposed here that aims at correctly modeling the behaviour of the human auditory system with respect to the direction-dependent localization blur and the different importance of different frequencies for localization. The localization blur of the human auditory system is a function of the source direction of the sound in the horizontal plane: It ranges from just a few degrees in the front up to 10 degrees to the side [Bla97]. It can be shown that the density of the coincidence units is similar to a Gaussian shape [SMH92]. This physiological finding can either be integrated into the model by a non-uniform chain of delay elements (cf. Figure 5.2) or by a weighting of the inhibited cross-correlation function depending on the internal delay index m. Since an additional weighting depending on the frequency is presented in the following, the integration by means of a weighting can be done with very little additional complexity.

These enhancements can be integrated into Equation 5.1 as follows:

$$\Psi\left(k,m\right)=\sum_{\mu=1}^{36}\sum_{i=k-N+1}^{k}\varphi_{n_l n_r}(\lambda,\mu)\mathrm{e}^{\frac{i-k}{N}}\cdot q\left(m,f_m(\mu)\right) \tag{5.3}$$

The overall weighting function therein is calculated as a multiplication of the two individual weightings which are presented in the following:

$$q(m,f)=q_1(m)\cdot q_2(f) \tag{5.4}$$

The shape of the first weighting factor $q_1(m)$ can be derived from [SMH92] as

$$q_1(m)=\frac{5}{3\cdot\sqrt{2\pi}}\cdot e^{-\frac{25}{18}\cdot\left(\frac{m\cdot\Delta\tau}{\mathrm{ms}}\right)^2}. \tag{5.5}$$

It has to be mentioned that both this internal time weighting and Gaik's proposal lead to more centrality, i.e., in situations where sources of similar intensity are active at the same time, the models will favor the centermost source. However, this is consistent with the localization blur of the auditory system.

A further improvement can be achieved using a frequency weighting taking into account which frequencies have more significant contributions to the localization accuracy. Raatgever [Raa80] has shown that especially a dominant region of frequencies around 600 Hz is more important for localizing. From this fact, a weighting function $q_2(f)$ can be derived to weight the contributions from different subbands [SZT88]:

$$q_2(f)=\begin{cases}10^{-\left(\sum_{i=1}^{3}b_i\cdot f^i\right)/10} & f<1200\,\mathrm{Hz}\\ 10^{-\left(\sum_{i=1}^{3}b_i\cdot 1200^i\right)/10} & f\geq 1200\,\mathrm{Hz}\end{cases} \tag{5.6}$$

with f in Hz and the coefficients of the polynomial as given in Table 5.1.

b_1	b_2	b_3
$-9.383\cdot10^{-2}$	$1.126\cdot10^{-4}$	$-3.992\cdot10^{-8}$

Table 5.1: Coefficients for frequency dependent weighting

The resulting overall weighting function $q(m,f)$ with the center frequency $f_m(\mu)$ of the frequency band μ is depicted in Figure 5.4. Looking at the internal delay axis, the Gaussian shape can be observed while $q(m,f)$ exhibits a parabolic shape along the frequency axis at low frequencies up to 1200 Hz and is constant for higher frequencies.

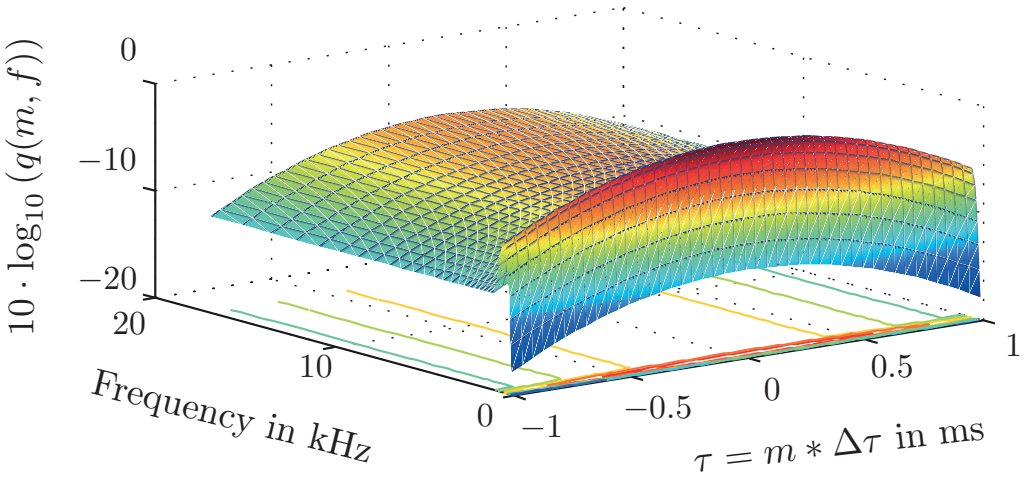


Figure 5.4: Combined weighting function $q(m, f)$ for the inhibited cross correlation

5.3.3 Evaluation

The use of the aforementioned enhancements leads to more reasonable *Neural Activity Patterns* (NAPs). The impact of the improvements to the binaural model is illustrated by means of the resulting binaural excitation patterns for a complex sound event, without (Figure 5.5) and with (Figure 5.6) the proposed model extensions. The example consists of a real recording of two male English speakers, one stationary at an angle of 30 ° on the left and one moving from 60 ° on the right to 0 °. The recording was carried out in a room with a reverberation time T_{60} of approximately 320 ms.

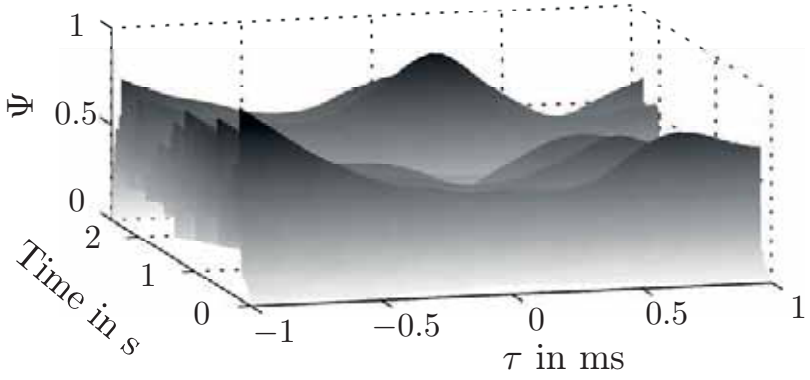


Figure 5.5: Binaural excitation pattern (original model)

When using the original model, a lot of the activities can be seen in the area of the monaural detectors (i.e., at values for τ of ± 1 ms). Note that even though there is no simple relation between the τ-axis and the source direction, values for τ between ± 0.6 ms approximately represent natural source directions while internal delays outside of this range indicate monaural signals.

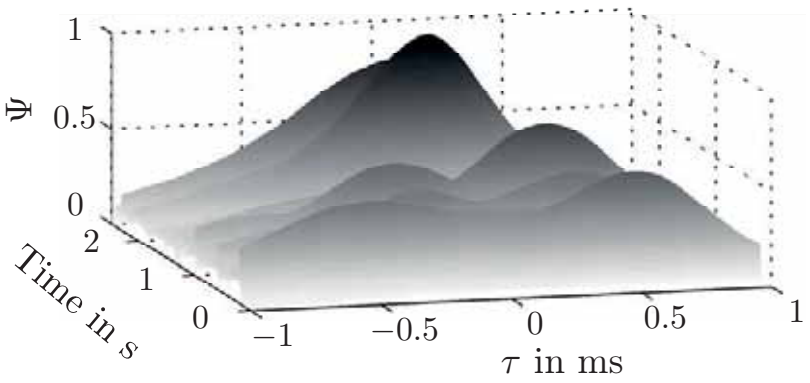


Figure 5.6: Binaural excitation pattern (improved model)

With the proposed improvements in place, the shape of the binaural excitation pattern is altered clearly and natural source directions are emphasized. As an additional pre-processing step for the blind clustering which will be described in more detail in Section 5.4, the resulting binaural excitation pattern is subject to a maximum search in every frame of 5 ms. The result of the maximum search is a two-dimensional distribution of maxima, the so-called *skeleton cross-correlogram*. The output of this step in this example is depicted in Figure 5.7.

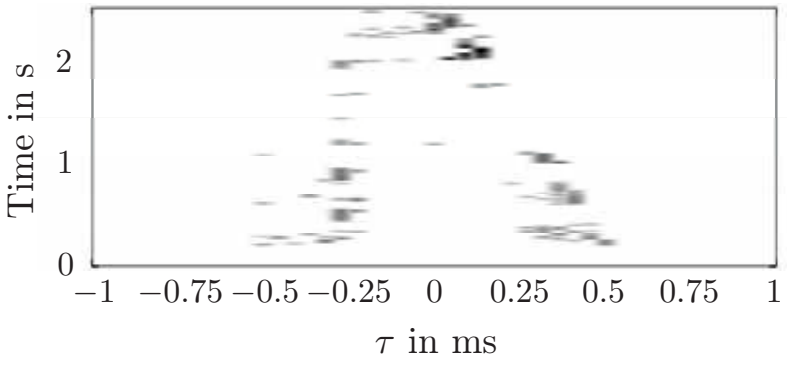


Figure 5.7: Skeleton cross-correlogram (improved model)

This *skeleton cross-correlogram* is the final output of the binaural hearing model and forms a good basis for the clustering process that is described in the next section. The hearing model belongs to the class of coincidence-based models and features several inclusions and improvements aiming at a more realistic and robust calculation of the *Neural Activity Patterns* (NAPs): an efficient weighting emphasizes the parts of the cross-correlogram which are most important for spatial perception and a maximum search removes the less important parts of the cross-correlogram.

5.4 Blind Clustering

Looking at Figure 5.7, it is obvious that there is one stationary source at $\tau \approx$ -0.25 ms as can be seen by the vertical line of maxima there. Simultaneously, a moving source is visible as the diagonal line from $\tau \approx 0.5$ ms to $\tau \approx 0$ ms in the diagram. Based on the improved hearing model, a novel cognitive processing scheme is proposed which can be used to estimate both the number of active sources as well as their direction.

5.4.1 Concept and Algorithm

The human brain is capable of separating different sources by identifying groups within the NAPs. The cognitive processing works in a very similar manner by applying *k-means clustering* to the skeleton cross-correlogram. The clustering uses the Euclidean distance and is described in more detail in [Bis95]. The k-means algorithm has, however, two drawbacks:

- The number of clusters needs to be known a priori.
- The clustering procedure is sensitive to initial centroids.

Both of these drawbacks can be controlled by a suitable initialization of the clustering process and a novel refinement step to improve the estimate for the number of active sources.

The distribution of the peaks in the skeleton cross-correlogram is depicted in the upper part of Figure 5.8. Each circle represents one maximum from Figure 5.7 that is above a threshold of 20% of the amplitude of the highest maximum. Below this plot, a histogram of the distributions of τ is shown. From this histogram, the initialization of the clustering algorithm can be done: The number of local maxima in the histogram is chosen as the number of clusters while the positions of the maxima are chosen as the initial centroids for the clusters.

In reverberant environments or for moving sources, this initialization can overestimate the number of active sources due to the fact that usually multiple local maxima appear in the histogram. In the presented example, three local maxima are found leading to the assumption that three sources are active. The k-means clustering for the example (cf. Figure 5.11) groups all the points on the left side ($-0.4\,\text{ms} < \tau < -0.1\,\text{ms}$ during the entire signal) into one cluster, the second cluster contains all the points in the top middle part of the figure

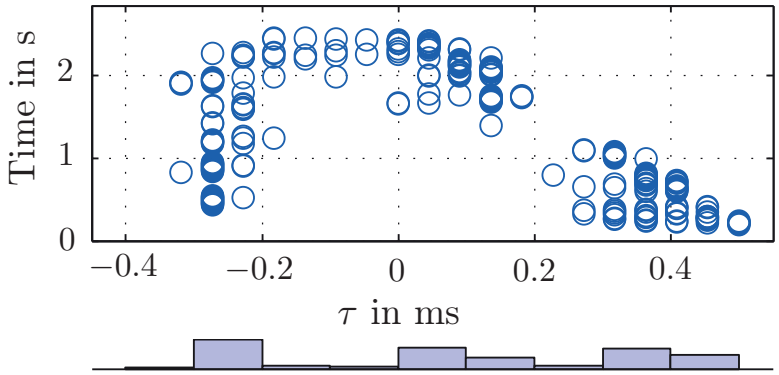


Figure 5.8: Skeleton cross-correlogram and derived histogram

($-0.1\,\text{ms} < \tau < 0.2\,\text{ms}$ for Time $> 1.5\,\text{s}$) and the remaining points in the lower right part form the third cluster ($0.2\,\text{ms} < \tau < 0.5\,\text{ms}$ for Time $< 1.5\,\text{s}$).

However, the second and the third cluster actually belong to one moving source. In order to accurately resolve this problem, a new refinement step is presented in Figure 5.9. Based on the initial k-means clustering, temporal and spectral information is extracted.

The temporal information consists of indicators about whether there is simultaneous or only successive neural activity in the three clusters. This information is derived from a comparison of the temporal spread of the clusters in the skeleton cross-correlogram. In the example, the first cluster is temporally overlapping with both other clusters which makes it unlikely that it has to be merged with one of the two. In contrast to that, the second and third cluster do not overlap which makes it possible that they belong to the same source.

In order to compare the spectral properties of individual clusters, a simple source separation is carried out. This is done for each cluster by combining all frames that belong to this cluster into one time domain signal. After that, spectral analyses of these time domain signals are performed by a single-level *Discrete Wavelet Transform* (DWT) with a Daubechies wavelet of order 6. This parameterization allows to discriminate even fairly similar signals, such as the two male speakers in the example.

The clusters are examined for similarities by calculating the zero-lag cross-correlation coefficients between the outputs of the DWT for the clusters. A statistical test is then applied which determines the probability of observing the calculated correlation coefficients by chance. Hence, smaller probabilities indicate more similarity between the clusters w.r.t. their spectral properties. The threshold for statistical significance was heuristically set to 10 % after numerous tests.

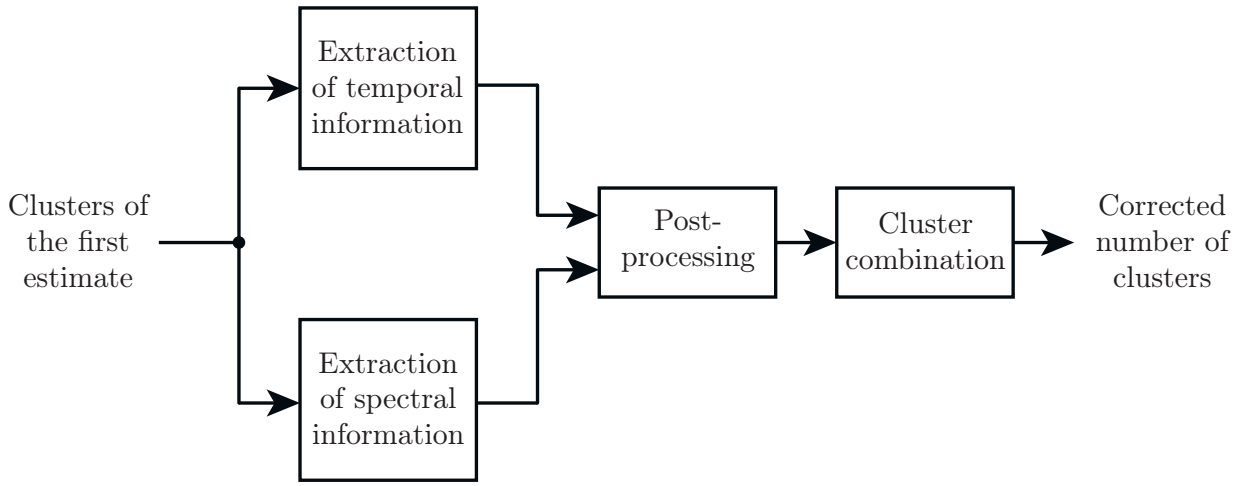


Figure 5.9: Refinement step for the estimation of the number of active sources

In the next step, the temporal and spectral informations are evaluated and every pair of clusters is assigned to one of three possible cases and the subsequent cluster combination acts according to the case that is found:

1. **The clusters are spectrally similar and are active simultaneously.**
   This is a case that occurs in strongly reverberant environments where the perceived width of a source is usually so large that it is registered as numerous sources coming from different directions in the first clustering. These clusters are combined unless the sources are spatially separated by more than 25 ° which is tested by comparing the clusters in the τ domain where an angle of 25 ° is approximately represented by a difference of 220 ms.

2. **The clusters are spectrally similar and are active sequentially.**
   This case happens when sources are active from time to time and silent in between (e.g., in a discussion with multiple participants). These clusters are always combined. Note that the test for temporal separation is less complex than the test for spatial separation making it worthwhile doing the test for temporal separation for all clusters and the test for spatial separation only for the clusters belonging to case 1.

3. **The clusters are spectrally dissimilar.**
   These clusters are not combined.

The corrected number of clusters allows to combine clusters as needed and to restart the clustering process in order to improve the initial clustering performance. This will be illustrated by an example in the next section.

5.4.2 Experimental Results and Limitations

The output of the initial clustering can be seen in Figure 5.10 where the moving source is incorrectly identified as two separate sources.

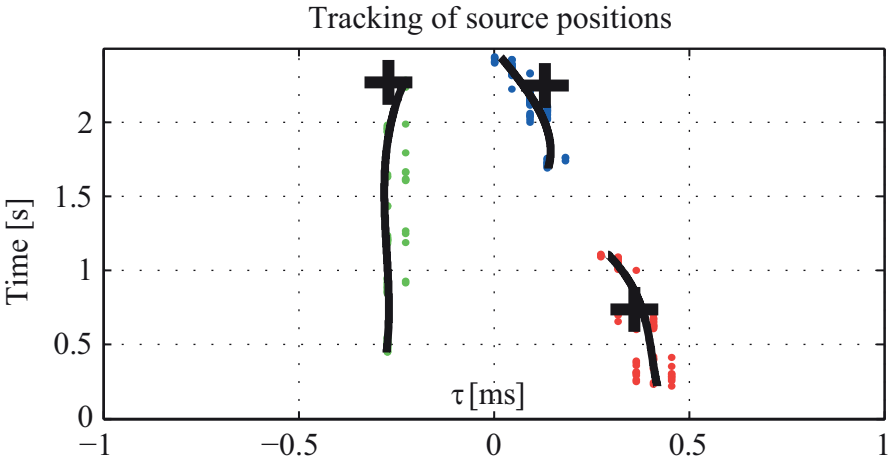


Figure 5.10: Initial clustering

With the refinement step, these two sources are joined to one as depicted in Figure 5.11 while the second source, also a male English speaker, is still correctly identified as another source.

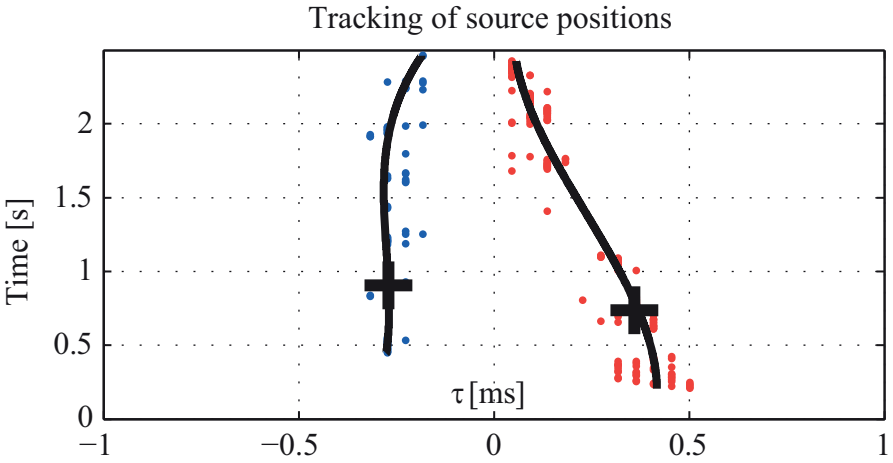


Figure 5.11: Initial and final clustering

While the model performs very well under various conditions, some limitations have to be mentioned. The robustness of the clustering decreases in very reverberant environments and for sources that are closer than 10 ° to each other. Additionally, since the model has neither the help of visual information nor the movement of the head, it is not able to resolve the ambiguity on the cone of confusion [Bla97]; it can only provide information about the azimuth angle and not about the elevation of the position of the source(s).

5.5 Extended PEAQ Measure for Binaural Signals

On the basis of the presented improved binaural hearing model, an *add-on* to the PEAQ model is proposed in the following which utilizes five additional parameters to represent the spatial properties of the signals that are derived from the presented binaural model of Section 5.3 and the clustering approach of Section 5.4. This setup allows to exploit the capabilities of PEAQ while simultaneously improving the performance for cases in which the spatial properties of the signals have a significant impact on the perceived quality. The approach is illustrated schematically in Figure 5.12. As can be seen from the figure, the non-linear mapping of the PEAQ output *Objective Difference Grade* (ODG) and the spatial parameters onto an overall quality measure is realized by means of a trained NN which is derived using an extensive listening test as reference. The result of the newly developed measure will be denoted as AODG in the following.

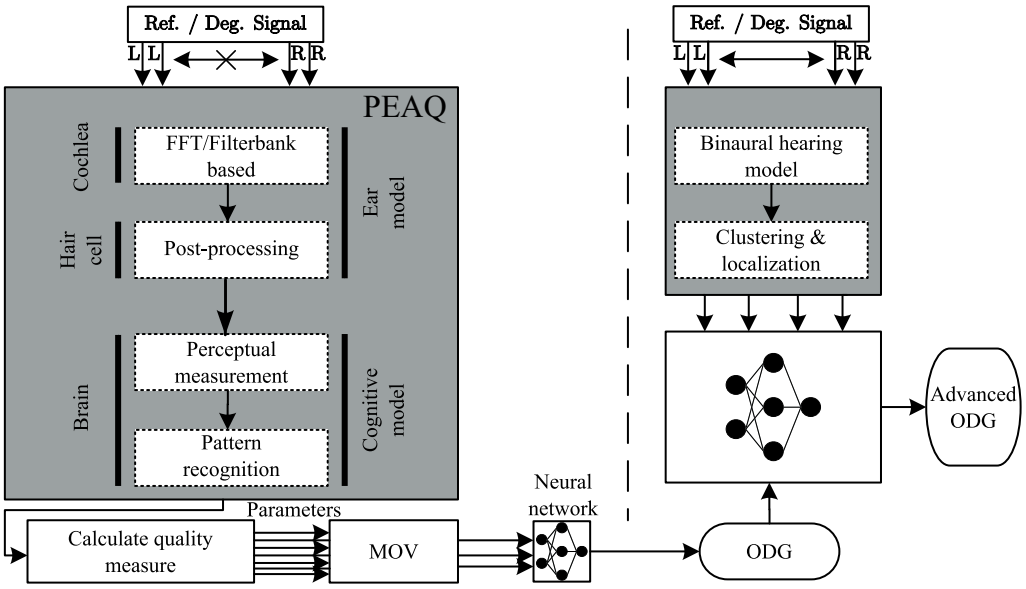


Figure 5.12: System overview of the novel quality measure

5.5.1 Spatial Quality Parameters

While many parameters can be calculated based on the hearing and cognitive model, most of the spatial information can already be extracted from the signals by a few well-chosen parameters. Having too many parameters also increases

the risk of overfitting the quality measure to the available training data which is always limited due to practical constraints.

In addition to the direct output of the hearing model, some parameters which can be derived from the clustering process are used here. Local maxima in the two-dimensional correlogram are grouped by a k-means clustering algorithm to identify clusters (i.e., sources). The clusters consist of multiple points in the k-m-plane grouped around the centroid $\mu_i(k,m)$ of the cluster i. Both an estimate Q_i of the spatial position of the source and a regression curve $R_i(k)$ are calculated for every cluster. The regression curve is a representation of the movement of the source over the length of the signal (cf. Figures 5.10 and 5.11).

Through analysis of a smaller, preliminary listening test, five parameters could be identified which provide a good representation of the spatial properties (with $\mathsf{E}_i\{\cdot\}$ denoting the expectation of $\cdot$ with respect to i):

- **Mean difference of correlograms:** The average difference between the correlograms of the reference signal and the degraded signal is determined according to:

$$p_1 = \mathsf{E}_k\left\{\mathsf{E}_m\left\{\left|\Psi_{\mathrm{ref}}\left(k,m\right) - \Psi_{\mathrm{deg}}\left(k,m\right)\right|\right\}\right\} \tag{5.7}$$

- **Mean difference of regression curves:** This parameter is calculated as the average of the absolute values of the difference between the regression curves for sources in the reference signal and regression curves for sources in the degraded signal.

$$p_2 = \mathsf{E}_i\left\{\mathsf{E}_k\left\{\left|R_{i,\mathrm{ref}}(k) - R_{i,\mathrm{deg}}(k)\right|\right\}\right\} \tag{5.8}$$

- **Mean difference of estimated source positions:** The average of the absolute values of all the differences between the estimated spatial source positions in the reference signal and the degraded signal is used as the third spatial parameter:

$$p_3 = \mathsf{E}_i\left\{\left|Q_{i,\mathrm{ref}} - Q_{i,\mathrm{deg}}\right|\right\} \tag{5.9}$$

- **Average difference of cluster centroids:** The average of the absolute values of the differences between the cluster centroids of the reference signal and the degraded signal is calculated as follows:

$$p_4 = \mathsf{E}_k\left\{\mathsf{E}_m\left\{\left|\mu_{i,\mathrm{ref}}(k,m) - \mu_{i,\mathrm{deg}}(k,m)\right|\right\}\right\} \tag{5.10}$$

- **Difference between the widths of the auditory events:** This parameter takes the difference in the width of the sources between the reference signal and the degraded signal into account. The width B_i of a source is determined from the internal delays τ_i that belong to this source (i.e., cluster i) as follows:

$$B_i = |\max(\tau_i) - \min(\tau_i)| \tag{5.11}$$

  The final spatial parameter is determined as the average of the changes in width of the sources:

$$p_5 = \mathsf{E}_i \{B_{i,\mathrm{ref}} - B_{i,\mathrm{deg}}\} \tag{5.12}$$

From these parameters, the spatial degradation in comparison to the reference signal can be measured instrumentally. Increasing values for these parameters indicate quality degradation.

5.6 Mapping Parameters to Advanced Objective Difference Grade

Even though the target of the proposed method is to avoid listening tests within the development process, it is of great importance that instrumental measures correctly include human perception. Hence they are trained based on the results of a suitable listening test.

5.6.1 Design of the Listening Test

The main focus during the development of AODG was on the overall audio quality while specifically taking degradation with respect to the spatial signal properties into account. A listening test as recommended in [ITU96a] was conducted which is described in the following.

The test used in this development is a *Degradation Category Rating* (DCR) test. In this test type, every participant gets to hear two signals:

- a reference signal of high quality and
- a degraded signal.

It is known to the participant which signal is the reference and the degraded signal, respectively. The test material is composed of speech and music signals containing fixed as well as moving sources. Different types of degradations (e.g., various codecs or a complete removal of all spatial properties by downmixing) were used to generate a meaningful set of test items.

The rating scale for this test consists of five rating levels according to [ITU96a] which can be found in Table 5.2.

Rating level r_{DCR}	Degradation is
5	inaudible
4	audible but not annoying
3	slightly annoying
2	annoying
1	very annoying

Table 5.2: Rating scale for the DCR test

Since standard PEAQ utilizes a rating scale of 0 to -4 for the ODG values, the rating levels are adjusted by $r = r_{\mathrm{DCR}} - 5$.

The listening test was conducted using the software tool *LIStening Test ENvironment* (LisTEn) [SSGV11]. The test took place in a quiet studio booth with very little reverberation (reverberation time $T_{60} = 120\,\mathrm{ms}$). The test signals were reproduced by a calibrated combination of a digital equalizer (Head Acoustics PEQ V) and a headphone (Sennheiser HD 600). In total, twenty listeners participated in the listening test that consisted of 50 test items per participant. A preliminary training phase with signals similar to the test signals was included before the test started.

5.6.2 Model Calibration

The crucial part in any instrumental quality measure is the mapping between parameters that are calculated from the audio signals and the quality estimate. This mapping is realized (as in standard PEAQ) by an NN which has a feed-forward structure with bias units on all layers, utilizes the spatial parameters p_1 to p_5 (cf. 5.5.1) and the output of standard PEAQ as its inputs, has one hidden layer consisting of ten neurons and a single output, the AODG value.

All neurons are connected by weighted edges so that every neuron can be characterized by its input and output edges and by its activation behaviour.

This activation behaviour is modeled for the hidden layer as a symmetrical sigmoid function on the sum x of the weighted input values:

$$\text{tansig}\,(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{5.13}$$

The reason for this choice is the fact that this function can be easily differentiated which is a necessary prerequisite for the applicability of many learning rules [VMR+88].

The NN needs to be trained first, this is done in a supervised manner by the Levenberg Marquardt method which is a very efficient method for small networks that converges quickly. It is described in detail in [HM94]. The training process can be summarized by the following steps:

1. Initialize the weighting factors randomly.
2. Pick two thirds (i.e., 34 of the 50) signals from the listening test randomly and in random order. The results of the listening test were averaged over all participants before training.
3. Let the NN calculate the output for the current weighting factors.
4. Modify the weighting factors and bias units by means of the learning algorithm in order to minimize the mean squared error between the output of the NN and the results of the listening test.
5. Control the learning process by the associated validation algorithm continuously and stop the learning process if no further gains are to be expected. This control inherently also helps to minimize the risk of overfitting.

The 16 signals that were not chosen for training are used for a later evaluation of the performance of the model.

This training regime has a disadvantage that has to be mentioned: The *best* Neural Network can only be determined after all $\binom{50}{34} = 4.92 \cdot 10^{12}$ possible selections of training signals are tested. As a reasonable compromise between quality and complexity, 2000 different selections were used to train 2000 NNs. Out of these NNs, the network with the highest correlation with the results of the listening test and the lowest mean square error was chosen (the details about this NN can be found in Appendix C).

Overfitting is an issue that can arise when training neural networks with limited amounts of training data. Due to the maximum length of a listening test that

does not lead to major discomfort for the participants, the training data in this setup is limited to 50 signals of which a third can not be used for training since it is necessary for evaluation purposes. An additional criterion is introduced that is specifically tailored to this application and should help in minimizing the risk of overfitting: Already in the development of standard PEAQ, a certain confidence interval was specified to define the allowed deviation between estimated and true quality, cf. Figure 5.13. The distance between quality estimates that are outside of the confidence interval and the confidence interval itself shall be minimized as well.

All input parameters are normalized for the training process. The normalization factors for this normalization are determined based on more than 1300 simulations for different signals and different signal processing systems. The maximum value for every parameter was calculated along with the 90^{th} percentile ($Q_{.90}$). These values are collected in Table 5.3.

Parameter	Maximum	$Q_{.90}$
p_1	12.05	4.61
p_2	1.25	0.39
p_3	1.36	0.41
p_4	1.26	0.29
p_5	1.37	1.14

Table 5.3: Maximum and 90^{th} percentile of the input parameters

Every parameter is then normalized accordingly:

$$f(p_i) = \begin{cases} \frac{p_i}{Q_{i,.90}}, & |p_i| < Q_{i,.90} \\ \text{sign}(p_i) \cdot Q_{i,.90}, & |p_i| \geq Q_{i,.90} \end{cases} \tag{5.14}$$

The clipping to the 90^{th} percentile reduces the impact of outliers and thus leads to a model which is more robust.

After determining the best NN according to the presented targets, it can then be evaluated with the remaining 16 signals that were not used for training. This evaluation is done in the next section.

5.7 Evaluation of the Proposed Quality Measure

The comparison between the novel quality measure and PEAQ, the basis for the development, can be carried out by different criteria:

- The correlation ρ between the quality estimate and the results of the listening test.
- The mean square error RMSE of the estimation compared to the results of the listening test.
- The coefficient of determination R^2 which is a measure for the ability of the model to generalize and approximate the true relationship between the input parameters and the estimated quality. Possible values for this measure are between 0 and 1, with higher values indicating a higher quality. The coefficient of determination is calculated from the results $r(i)$ of the listening test, their average $\bar{r}(i)$, and the quality estimates $\hat{r}(i)$ as

$$R^2 = 1 - \frac{\sum_{i=1}^{50} \big(r(i) - \hat{r}(i)\big)^2}{\sum_{i=1}^{50} \big(r(i) - \bar{r}(i)\big)^2} \,. \tag{5.15}$$

- The number of outliers $N_{d_{out}}$ with respect to the previously defined confidence interval. It is worth noting that a model with fewer outliers is not necessarily the better model in all cases if these outliers are more severe (i.e. further off from the confidence interval).
- The rank correlation ρ_{rank} between the quality estimate and the results of the listening test.

With these quality measures, a comparison of PEAQ and the proposed instrumental measure can be carried out. This comparison is done with those signals that were not used for training the NN. In the diagrams, all 50 signals are depicted to get a better impression of the performance of PEAQ. In the diagram with the results of the new instrumental measure, the signals that were used for training are clearly marked to also allow for a quick overview on the performance for signals that were not included in the training process.

In Figures 5.13 and 5.14, the results when using standard PEAQ or the proposed instrumental measure can be seen. An ideal instrumental quality measure would place all individual data points on (or very close to) the dashed main diagonal.

The positive conclusion that can be taken from Figure 5.13 is the fact that most of the markers are within the confidence interval and that there are no major outliers to the right of the confidence interval, i.e., there are no cases for which PEAQ strongly overestimates the quality of the signals. On the other hand, there are numerous cases of strong underestimation of the signal quality which can be seen from the various points in the top left part of the diagram.

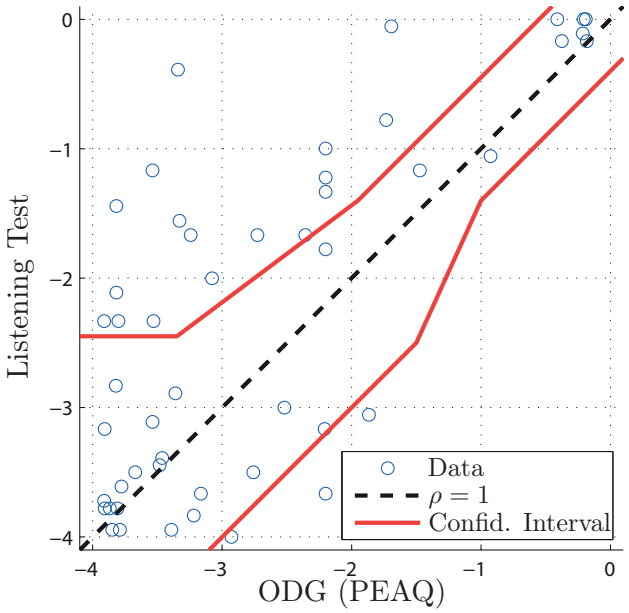


Figure 5.13: Scatter plot of listening test results and ODG values according to PEAQ

It can clearly be seen from the results for the proposed instrumental measure in Figure 5.14 that explicitly including the spatial parameters leads to a significantly stronger correlation between the results of the listening test and the quality estimates of the instrumental measure. The number of outliers outside

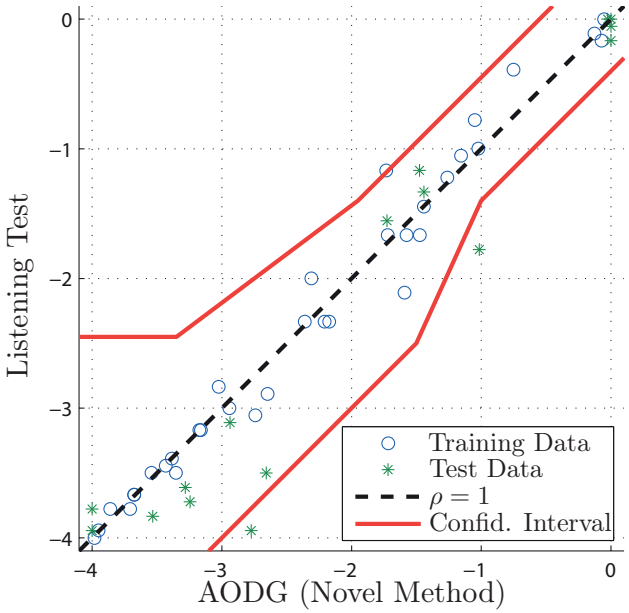


Figure 5.14: Scatter plot of listening test results and AODG values according to the proposed quality measure

of the confidence interval is very small and even these outliers are very close to the confidence intervals.

As a more formal and clearer comparison of the performance of both standard PEAQ and the proposed instrumental measure, the aforementioned criteria are calculated for both measures and collected in Table 5.4.

Criterion \ Measure	ODG	AODG	AODG (Test Data Only)
ρ	0.704	0.971	0.954
RMSE	1.067	0.322	0.497
R^2	0.179	0.942	0.898
$N_{d_{out}}$	40%	6%	12%
ρ_{rank}	0.626	0.950	0.881

Table 5.4: Comparison between PEAQ and the proposed instrumental measure

All criteria illustrate the improved performance of the proposed instrumental measure compared to standard PEAQ. Especially the column of values that are calculated only for the signals that were not used for training the model is important for quantifying the performance of AODG. These values and the high coefficient of determination clearly indicate that the measure will perform accordingly for other test cases and signals.

As an example, a binaural piece of music (consisting of a piano, a violin and a trumpet playing from different directions) is coded by *MPEG-1 layer 3* (MP3) [BS96] and parametric stereo [BvdPKS05]. Both systems are used at configurations that will not lead to a really good transmission quality, but the decisive difference between the two coding systems is that parametric stereo explicitly transmits and reconstructs spatial parameters. This difference leads to a more natural and consistent spatial impression. This example was also included in the listening test, where the quality for the transmission with the MP3 system in this configuration was identified as quite bad while the transmission with parametric stereo is still acceptable for this signal. The results for both transmission systems and both instrumental quality measures are collected in Table 5.5 along with the results of the listening test.

It is obvious that while both standard PEAQ and the AODG correctly indicate that the transmission quality of the MP3 system in this configuration is not good at all, only AODG is able to correctly identify and differentiate the improved subjective quality of the transmission with parametric stereo.

Measure \ Transmission system	MP3	parametric stereo
ODG	−3.5	−3.8
AODG	−2.9	−1.4
Listening Test	−3.1	−1.4

Table 5.5: Instrumental quality measures and listening test results for different transmission systems

5.8 Conclusions

An advanced binaural auditory model and a mathematical model of cognitive behavior was proposed and shown to be able of estimating both the position and the number of acoustic sources. The model works reliably even for multiple concurrent speakers and moving targets. Possible application scenarios for the modeling scheme include all signal processing schemes that rely on an accurate representation of the human hearing system as well as source separation algorithms that require knowledge about the number of active sources.

An extension to PEAQ was presented which makes use of five parameters that are derived from the presented binaural hearing model. These parameters provide a compact description of the signal properties that are important for spatial perception. The extension follows the basic principle of PEAQ by calculating these parameters and then mapping them onto the quality measure by means of a Neural Network. The inclusion of spatial information into the instrumental quality measurement leads, in contrast to PEAQ, to a consistently high correlation between the instrumental measure and a listening test for stereo signals.

6

Summary

Throughout this thesis, different aspects of multi channel signal processing have been considered and novel analyses and algorithms for different parts of a multi channel audio transmission system were presented. The results provide theoretical performance bounds for parts of the transmission system as well as practical algorithms for both enhancement and transmission of multi channel signals.

Most of the presented results utilize a two stage system that contains novel methods for exploiting and transmitting spatial properties of audio signals. Three applications for the outer stage are given: On the transmitting side, a beamforming algorithm or a mixing of the microphone signals can be derived from the generic system. On the receiving side, the counterpart to the mixing process from the transmitting side or a method to improve speech intelligibility by a postfiltering procedure are found in this stage. The inner stage is a multi channel predictive coding scheme that was shown to allow for an efficient transmission. Noise shaping concepts for multi channel predictive systems received major attention in this part since no suitable concepts were known for this task so far.

Beginning at acoustic front end on the transmitting side, a beamforming algorithm was presented in Section 3.3. This algorithm belongs to the class of filter-and-sum beamformers and is specifically tailored for use with sources in the near field. A simulation of the reception characteristic forms the basis of the determination of the filter coefficients. This reception characteristic is then compared with a target for the reception characteristic and a numerical optimization algorithm is applied to approximate the target as closely as possible. This procedure offers two main advantages: It is very flexible with respect to the area in which the beamformer shall be used. It can easily be configured for arbitrary microphone positionings and it allows to optimize for the near field

or for the far field by simply changing the impulse responses that are used. Additionally, it explicitly takes the actual acoustic environment into account when being used with measured impulse responses.

The performance of the algorithm was evaluated and a comparison with other beamforming concepts confirmed that it achieves a better approximation of the target reception characteristic. The beamforming algorithm was developed in the context of a video conferencing application where it was implemented for a real time communication system so that its performance was also verified under realistic conditions.

The postfiltering for an increased speech intelligibility was presented in Section 3.4. It utilizes knowledge from room acoustics about the positive influence of certain room impulse responses on the speech intelligibility to design a postfilter for single or multi channel signal transmission systems. A thorough evaluation of different types of room impulse response models revealed that a very simple model for early reflections leads to the best performance. This sparse impulse response model was shown to consistently increase the speech intelligibility (as measured by the *Speech Transmission Index* (STI)) for speech transmission with the *Adaptive Multi-Rate Wideband* (AMR-WB) codec.

The other use case for the outer stage, the preconditioning, is closely related to the inner stage. The inner stage exploits the symmetries that stem from the mixing that has taken place in the outer stage.

The multi channel prediction scheme that forms the inner stage allows for a hierarchical extension of existing single channel transmission systems with very little additional algorithmic delay. In the area of multi channel prediction systems, several aspects have been considered in this thesis. A new analysis of the achievable prediction gains in a joint inter and intra channel prediction setup in comparison to different sequential inter and intra channel prediction setups was presented in Section 2.5 and exemplified based on several small scale examples. Two alternative paradigms for adaptively distributing the available filter taps between inter and intra channel prediction were devised in Section 2.6. Fundamentally, they are closely related to Long Term Prediction and a derived very simple coding scheme is shown to achieve significantly larger prediction gains than common inter and intra channel prediction systems.

In Chapter 4, a realization of a flexible multi channel prediction system is presented and analyzed. Noise shaping concepts for multi channel predictive systems are introduced that are shown to achieve different effective quantization noise signals at the output which, e.g., allow to have identical transmission qualities in all channels. The impact of the system dimensions with respect to

the length of the prediction filters or the word length of the used quantizers is evaluated based on *Perceptual Evaluation of Audio Quality* (PEAQ) and *Advanced Objective Difference Grade* (AODG). It can be deduced that the data rate efficiency for all variants of the proposed system is significantly higher than for an independent transmission of the input channels which uses the same quantizers. A combination of single channel open loop predictors with the novel noise shaping concept shows the best results for both instrumental quality measures.

One of these instrumental measures, the *Advanced Objective Difference Grade* (AODG), is introduced in Chapter 5. Its foundation is an advanced binaural auditory model and a mathematical modeling of cognitive behavior that were proposed and shown to be able of estimating both the position and the number of acoustic sources. The model works reliably even for multiple concurrent speakers and moving targets. Possible application scenarios for the modeling scheme include all signal processing schemes that rely on an accurate representation of the human hearing system as well as source separation algorithms that require knowledge about the number of active sources.

AODG was designed as an extension to PEAQ. It makes use of five parameters that are derived from the aforementioned hearing and cognitive model. These parameters provide a compact description of the signal properties that are important for spatial perception. The extension follows the basic principle of PEAQ by calculating these parameters and then mapping them onto the quality measure by means of a Neural Network. The inclusion of spatial information into the instrumental quality measurement leads, in contrast to PEAQ, to a consistently high correlation between the instrumental measure and a listening test.

A

Prediction Errors: Joint and Sequential Optimization

In Section 2.5.3, only the final expressions for the energies of the prediction error signals are given in Equations 2.46, 2.49 and 2.52 for brevity. The complete derivations can be found here.

A.1 Sequential Optimization – Intra Channel Prediction First

The prediction error $z_1(k)$ in this case can be written as

$$z_1(k) = y_1(k) - h_{\text{intra}}(1) \cdot y_1(k-1) - h_{\text{inter}}(0) \cdot y_c(k)\,. \tag{A.1}$$

With the two filter coefficients

$$h_{\text{intra}}(1) = \frac{\varphi_{y_1 y_1}(1)}{\varphi_{y_1 y_1}(0)} \tag{A.2}$$

and

$$h_{\text{inter}}(0) = \frac{\varphi_{y_1 y_1}(0) \cdot \varphi_{y_1 y_c}(0) - \varphi_{y_1 y_1}(1) \cdot \varphi_{y_1 y_c}(1)}{\varphi_{y_1 y_1}(0) \cdot \varphi_{y_c y_c}(0)}\,, \tag{A.3}$$

the energy of the prediction error can be calculated:

$$
\begin{aligned}
\mathrm{E}\left\{z_1^2(k)\right\} =& \mathrm{E}\left\{\left(y_1(k) - h_{\text{intra}}\left(1\right)\cdot y_1(k-1) - h_{\text{inter}}\left(0\right)\cdot y_c(k)\right)^2\right\}\\
=& \mathrm{E}\Big\{y_1^2(k) + h_{\text{intra}}^2(1)\cdot y_1^2(k-1) + h_{\text{inter}}^2(0)\cdot y_c^2(k)\\
& -2y_1(k)\cdot h_{\text{intra}}\left(1\right)\cdot y_1(k-1) - 2y_1(k)\cdot h_{\text{inter}}\left(0\right)\cdot y_c(k)\\
& +2h_{\text{intra}}\left(1\right)\cdot y_1(k-1)\cdot h_{\text{inter}}\left(0\right)\cdot y_c(k)\Big\}\\
=& \varphi_{y_1y_1}(0) + h_{\text{intra}}^2(1)\cdot\varphi_{y_1y_1}(0) + h_{\text{inter}}^2(0)\cdot\varphi_{y_cy_c}(0)\\
& - 2h_{\text{intra}}\left(1\right)\cdot\varphi_{y_1y_1}(1) - 2h_{\text{inter}}\left(0\right)\varphi_{y_1y_c}(0)\\
& + 2h_{\text{intra}}\left(1\right)\cdot h_{\text{inter}}\left(0\right)\cdot\varphi_{y_1y_c}(1) \qquad \text{(A.4)}\\
=& \varphi_{y_1y_1}(0) + \left(\frac{\varphi_{y_1y_1}(1)}{\varphi_{y_1y_1}(0)}\right)^2\cdot\varphi_{y_1y_1}(0) - 2\frac{\varphi_{y_1y_1}(1)}{\varphi_{y_1y_1}(0)}\cdot\varphi_{y_1y_1}(1)\\
& + \left(\frac{\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0)}\right)^2\cdot\varphi_{y_cy_c}(0)\\
& - 2\frac{\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0)}\cdot\varphi_{y_1y_c}(0)\\
& + 2\frac{\varphi_{y_1y_1}(1)}{\varphi_{y_1y_1}(0)}\cdot\frac{\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0)}\varphi_{y_1y_c}(1)\\
=& \frac{1}{\varphi_{y_1y_1}^2(0)\cdot\varphi_{y_cy_c}(0)}\cdot\Big(\varphi_{y_1y_1}^3(0)\cdot\varphi_{y_cy_c}(0)\\
& +\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_1}^2(1)\cdot\varphi_{y_cy_c}(0)\\
& - 2\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_1}^2(1)\cdot\varphi_{y_cy_c}(0)\\
& + \varphi_{y_1y_1}^2(0)\cdot\varphi_{y_1y_c}^2(0)\\
& - 2\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)\\
& + \varphi_{y_1y_1}^2(1)\cdot\varphi_{y_1y_c}^2(1)\\
& - 2\varphi_{y_1y_1}^2(0)\cdot\varphi_{y_1y_c}^2(0) + 2\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)\\
& + 2\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1) - 2\varphi_{y_1y_1}^2(1)\cdot\varphi_{y_1y_c}^2(1)\\
=& \frac{1}{\varphi_{y_1y_1}^2(0)\cdot\varphi_{y_cy_c}(0)}\cdot\Big(2\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)\\
& + \varphi_{y_1y_1}^3(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_1}^2(0)\cdot\varphi_{y_1y_c}^2(0) - \varphi_{y_1y_1}^2(1)\cdot\\
& \cdot\varphi_{y_1y_c}^2(1) - \varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_1}^2(1)\cdot\varphi_{y_cy_c}(0)\Big)\,. \qquad \text{(A.5)}
\end{aligned}
$$

A.2 Sequential Optimization – Inter Channel Prediction First

Here, the prediction error can be written as

$$z_1(k) = y_1(k) - h_{\text{intra}}(1) \cdot y_1(k-1) - h_{\text{inter}}(0) \cdot y_c(k) + h_{\text{intra}}(1) \cdot h_{\text{inter}}(0) \cdot y_c(k-1) \,. \tag{A.6}$$

Calculating the energy of the prediction error for this case gives:

$$\begin{aligned}
\mathrm{E}\left\{z_1^2(k)\right\} =& \mathrm{E}\left\{\Big(y_1(k) - h_{\text{intra}}(1) \cdot y_1(k-1) - h_{\text{inter}}(0) \cdot y_c(k) + \right. \\
& \left. h_{\text{intra}}(1) \cdot h_{\text{inter}}(0) \cdot y_c(k-1)\Big)^2\right\} \\
=& \mathrm{E}\Big\{y_1^2(k) - h_{\text{intra}}^2(1) \cdot y_1^2(k-1) - h_{\text{inter}}^2(0) \cdot y_c^2(k) \\
& + h_{\text{intra}}^2(1) \cdot h_{\text{inter}}^2(0) \cdot y_c^2(k-1) \\
& - 2y_1(k) \cdot h_{\text{intra}}(1) \cdot y_1(k-1) \\
& - 2y_1(k) \cdot h_{\text{inter}}(0) \cdot y_c(k) \\
& + 2y_1(k) \cdot h_{\text{intra}}(1) \cdot h_{\text{inter}}(0) \cdot y_c(k-1) \\
& + 2h_{\text{intra}}(1) \cdot h_{\text{inter}}(0) \cdot y_1(k-1) \cdot y_c(k) \\
& - 2h_{\text{intra}}^2(1) \cdot h_{\text{inter}}(0) \cdot y_1(k-1) \cdot y_c(k-1) \\
& - 2h_{\text{intra}}(1) \cdot h_{\text{inter}}^2(0) \cdot y_c(k) \cdot y_c(k-1)\Big\}
\end{aligned}$$

$$
\begin{aligned}
=&\varphi_{y_1y_1}(0) - h^2_{\text{intra}}(1)\cdot\varphi_{y_1y_1}(0) - h^2_{\text{inter}}(0)\cdot\varphi_{y_cy_c}(0)\\
&+ h^2_{\text{intra}}(1)\cdot h^2_{\text{inter}}(0)\cdot\varphi_{y_cy_c}(0)\\
&- 2h_{\text{intra}}(1)\cdot\varphi_{y_1y_1}(1)\\
&- 2h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(0)\\
&+ 2h_{\text{intra}}(1)\cdot h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(1)\\
&+ 2h_{\text{intra}}(1)\cdot h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(-1)\\
&- 2h^2_{\text{intra}}(1)\cdot h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(0)\\
&- 2h_{\text{intra}}(1)\cdot h^2_{\text{inter}}(0)\cdot\varphi_{y_cy_c}(0)\\
=&\left(1+h_{\text{intra}}(1)^2\right)\cdot\left(\varphi_{y_1y_1}(0) - 2h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(0)\right.\\
&\left.+h_{\text{inter}}(0)^2\cdot\varphi_{y_cy_c}(0)\right) + 2h_{\text{intra}}(1)\cdot\left(h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(1)\right.\\
&\left.+h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(-1) - \varphi_{y_1y_1}(1) - h_{\text{inter}}(0)^2\cdot\varphi_{y_cy_c}(1)\right)
\end{aligned}
\tag{A.7}
$$

A.3 Joint Optimization

The prediction error in this case is identical to the case in Section A.1

$$
z_1(k) = y_1(k) - h_{\text{intra}}(1)\cdot y_1(k-1) - h_{\text{inter}}(0)\cdot y_c(k) \tag{A.8}
$$

Due to the fact that the two prediction errors are identical, we can directly start at the intermediate result from Section A.1 that is labeled Equation A.4 and insert the filter coefficients

$$
h_{\text{intra}}(1) = \frac{\varphi_{y_1y_1}(1)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi^2_{y_1y_c}(1)} \tag{A.9}
$$

and

$$
h_{\text{inter}}(0) = \frac{\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi^2_{y_1y_c}(1)}\,. \tag{A.10}
$$

This results in the energy:

$$
\begin{aligned}
&\mathrm{E}\left\{z_1^2(k)\right\} = \varphi_{y_1y_1}(0) + h_{\text{intra}}^2(1)\cdot\varphi_{y_1y_1}(0) + h_{\text{inter}}^2(0)\cdot\varphi_{y_cy_c}(0)\\
&-2h_{\text{intra}}(1)\cdot\varphi_{y_1y_1}(1) - 2h_{\text{inter}}(0)\,\varphi_{y_1y_c}(0)\\
&+2h_{\text{intra}}(1)\cdot h_{\text{inter}}(0)\cdot\varphi_{y_1y_c}(1)\\
&=\varphi_{y_1y_1}(0) + \left(\frac{\varphi_{y_1y_1}(1)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)}\right)^2\cdot\varphi_{y_1y_1}(0)\\
&+\left(\frac{\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)}\right)^2\cdot\varphi_{y_cy_c}(0)\\
&-2\frac{\varphi_{y_1y_1}(1)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)}\cdot\varphi_{y_1y_1}(1)\\
&-2\frac{\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)}\varphi_{y_1y_c}(0)\\
&+2\frac{\varphi_{y_1y_1}(1)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)}\\
&\cdot\frac{\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)}{\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)}\cdot\varphi_{y_1y_c}(1)\\
&=\frac{1}{\left(\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)\right)^2}\\
&\cdot\Big(\varphi_{y_1y_1}(0)\cdot\left(\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)\right)^2\\
&-2\left(\varphi_{y_1y_1}^2(1)\cdot\varphi_{y_cy_c}^2(1) - \varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)\cdot\varphi_{y_1y_1}(1)\right.\\
&\left.+\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}^2(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)\right)\\
&\cdot\left(\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}^2(1)\right)\\
&+\left(\varphi_{y_1y_1}(1)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)\right)^2\cdot\varphi_{y_1y_1}(0)\\
&+\left(\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)\right)^2\cdot\varphi_{y_cy_c}(0)\\
&+2\left(\varphi_{y_1y_1}(1)\cdot\varphi_{y_cy_c}(0) - \varphi_{y_1y_c}(0)\cdot\varphi_{y_1y_c}(1)\right)\\
&\cdot\left(\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0) - \varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(1)\right)\cdot\varphi_{y_1y_c}(1)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\left(\varphi_{y_1y_1}(0)\cdot\varphi_{y_cy_c}(0)-\varphi^2_{y_1y_c}(1)\right)^2}\cdot\Big(\varphi^3_{y_1y_1}(0)\cdot\varphi^2_{y_cy_c}(0)\\
&-\varphi^2_{y_1y_1}(0)\cdot\varphi^2_{y_1y_c}(0)\cdot\varphi_{y_cy_c}(0)-\varphi_{y_1y_1}(0)\cdot\varphi^2_{y_1y_1}(1)\cdot\varphi^2_{y_cy_c}(0)\\
&+\varphi_{y_1y_1}(0)\cdot\varphi^4_{y_1y_c}(1)-2\varphi^2_{y_1y_1}(0)\cdot\varphi^2_{y_1y_c}(1)\cdot\varphi_{y_cy_c}(0)\\
&+\varphi_{y_1y_1}(0)\cdot\varphi^2_{y_1y_c}(0)\cdot\varphi^2_{y_1y_c}(1)+3\varphi^2_{y_1y_1}(1)\cdot\varphi^2_{y_1y_c}(1)\cdot\varphi_{y_cy_c}(0)\\
&-2\varphi_{y_1y_1}(1)\cdot\varphi_{y_1y_c}(0)\cdot\varphi^3_{y_1y_c}(1)-2\varphi_{y_1y_1}(1)\cdot\varphi^3_{y_1y_c}(1)\cdot\varphi_{y_cy_c}(0)\\
&+2\,\varphi_{y_1y_1}(0)\cdot\varphi_{y_1y_c}(0)\cdot\varphi^2_{y_1y_c}(1)\cdot\varphi_{y_cy_c}(0)\Big)
\end{aligned}
\tag{A.11}
$$

B

Symmetries in the Predictive Coding Scheme

The symmetries in the vectors of filter coefficients and the prediction error signals (cf. Equations 4.7 and 4.8) were derived in [KV08] for a related system setup. In the original publication, there were constraints on the prediction filters – they were required to have a linear phase response. The derivation of the symmetry properties is given here for the case of unconstrained prediction filters.

The prediction errors $z_a(k)$ are given (cf. Equation 4.4) by

$$z_a(k) = y_c(k) - \sum_{\lambda=0}^{L-1} h_a(\lambda) \cdot z_{\mathrm{A}}\left(k-\lambda\right) . \tag{B.1}$$

As presented in Section 4.2, the filter coefficients can be determined by solving the set of equations

$$\mathbf{Z}_{\mathrm{AA}} \cdot \mathbf{h}_a = \mathbf{Y}_{c\mathrm{A}} . \tag{B.2}$$

As long as $\mathbf{Z}_{\mathrm{AA}}$ can be inverted, this can be rearranged to

$$\mathbf{h}_a = \mathbf{Z}_{\mathrm{AA}}^{-1} \cdot \mathbf{Y}_{c\mathrm{A}} . \tag{B.3}$$

The quadratic matrix $\mathbf{Z}_{\mathrm{AA}}$ therein contains the autocorrelation values $\varphi_{z_{\mathrm{A}}z_{\mathrm{A}}}(\lambda)$ of the main channel in symmetric Toeplitz structure:

$$\mathbf{Z}_{\mathrm{AA}} = \begin{pmatrix} \varphi_{z_{\mathrm{A}}z_{\mathrm{A}}}(0) & \cdots & \varphi_{z_{\mathrm{A}}z_{\mathrm{A}}}(L-1) \\ \varphi_{z_{\mathrm{A}}z_{\mathrm{A}}}(1) & \cdots & \varphi_{z_{\mathrm{A}}z_{\mathrm{A}}}(L-2) \\ \vdots & \ddots & \vdots \\ \varphi_{z_{\mathrm{A}}z_{\mathrm{A}}}(L-1) & \cdots & \varphi_{z_{\mathrm{A}}z_{\mathrm{A}}}(0) \end{pmatrix} \tag{B.4}$$

The column vector $\mathbf{h}_a$ is composed of the filter coefficients $h_a(k)$:

$$\mathbf{h}_a = \Big(h_a(0)\; h_a(1)\; \ldots\; h_a(L-1) \Big)^{\mathrm{T}} \tag{B.5}$$

The column vector $\mathbf{Y}_{c\mathrm{A}}$ contains the cross correlation values $\varphi_{y_c z_\mathrm{A}}(\lambda)$ between the respective input channel c and the main channel:

$$\mathbf{Y}_{c\mathrm{A}} = \Big(\varphi_{y_c z_\mathrm{A}}(0)\; \varphi_{y_c z_\mathrm{A}}(-1)\; \ldots\; \varphi_{y_c z_\mathrm{A}}(1-L) \Big)^{\mathrm{T}} \tag{B.6}$$

Calculating the sum of all vectors of filter coefficients gives:

$$\begin{aligned}
\sum_{a=1}^{\mathrm{A}-1} \mathbf{h}_a &= \sum_{a=1}^{\mathrm{A}-1} \mathbf{Z}_{\mathrm{AA}}^{-1} \cdot \mathbf{Y}_{c\mathrm{A}} \\
&= \mathbf{Z}_{\mathrm{AA}}^{-1} \cdot \sum_{a=1}^{\mathrm{A}-1} \mathbf{Y}_{c\mathrm{A}} \\
&= \mathbf{Z}_{\mathrm{AA}}^{-1} \cdot \mathbf{Y}_{\mathrm{AA}} \\
&= \big(\underbrace{0 \ldots 0}_{\tau} \quad 1 \quad \underbrace{0 \ldots 0}_{\tau} \big)^{\mathrm{T}}
\end{aligned} \tag{B.7}$$

With the column vector $\mathbf{Y}_{\mathrm{AA}}$ containing the autocorrelation values $\varphi_{y_\mathrm{A} z_\mathrm{A}}(\lambda)$ (note the index shift in comparison to $\mathbf{Y}_{c\mathrm{A}}$ due to the delays in the outer stage according to Equation 4.3):

$$\begin{aligned}
\mathbf{Y}_{\mathrm{AA}} &= \Big(\varphi_{z_\mathrm{A} z_\mathrm{A}}(\tfrac{L-1}{2})\; \ldots\; \varphi_{z_\mathrm{A} z_\mathrm{A}}(0)\; \ldots\; \varphi_{z_\mathrm{A} z_\mathrm{A}}(-\tfrac{L-1}{2}) \Big)^{\mathrm{T}} \\
&= \Big(\varphi_{z_\mathrm{A} z_\mathrm{A}}(\tfrac{L-1}{2})\; \ldots\; \varphi_{z_\mathrm{A} z_\mathrm{A}}(0)\; \ldots\; \varphi_{z_\mathrm{A} z_\mathrm{A}}(\tfrac{L-1}{2}) \Big)^{\mathrm{T}}
\end{aligned} \tag{B.8}$$

Using Equations B.7 and 4.3, the sum of all prediction errors can be calculated:

$$\begin{aligned}
\sum_{a=1}^{\mathrm{A}-1} z_a(k) &= \sum_{a=1}^{\mathrm{A}-1} \left(y_c(k) - \sum_{\lambda=0}^{L-1} h_a(\lambda) \cdot z_{\mathrm{A}}\left(k-\lambda\right) \right) \\
&= \sum_{a=1}^{\mathrm{A}-1} \left(y_c(k) - z_{\mathrm{A}}\left(k - \frac{L-1}{2} \right) \right) \\
&= \sum_{a=1}^{\mathrm{A}-1} y_c(k) - (\mathrm{A}-1) \cdot z_{\mathrm{A}}\left(k - \frac{L-1}{2} \right) \\
&= 0
\end{aligned} \tag{B.9}$$

C

Coefficients of the Neural Network

The *Neural Network* (NN) that is utilized in the novel quality measure to map the derived spatial and quality parameters to the final *Advanced Objective Difference Grade* (AODG) is described here in detail. The NN has seven inputs and a single output: The inputs are the five parameters that are described in Section 5.5.1 ($p_1 \dots p_5$), the *Objective Difference Grade* (ODG) value according to the *Perceptual Evaluation of Audio Quality* (PEAQ) measure (p_6) and an additional combination of p_1 and p_6 according to $p_7 = p_1 \cdot p_6$. The output o_1 is the AODG value for the tested audio signal.

A block diagram of the NN is depicted in Figure C.1. The row input vector $\mathbf{p}$ consists of the seven input parameters p_i and is connected by the weight matrix $\mathbf{w}_1$ to one hidden layer with ten neurons n_i (hidden layer vector $\mathbf{n} = [n_1 \dots n_{10}]$). This hidden layer is then connected by the weight vector $\mathbf{w}_2$ to the output o_1. After the input and after the hidden layer, there are also bias units ($\mathbf{b}_1$ between input and hidden layer and b_2 between hidden layer and output, respectively) and the ten neurons of the hidden layer utilize a sigmoid characteristic as their activation function.

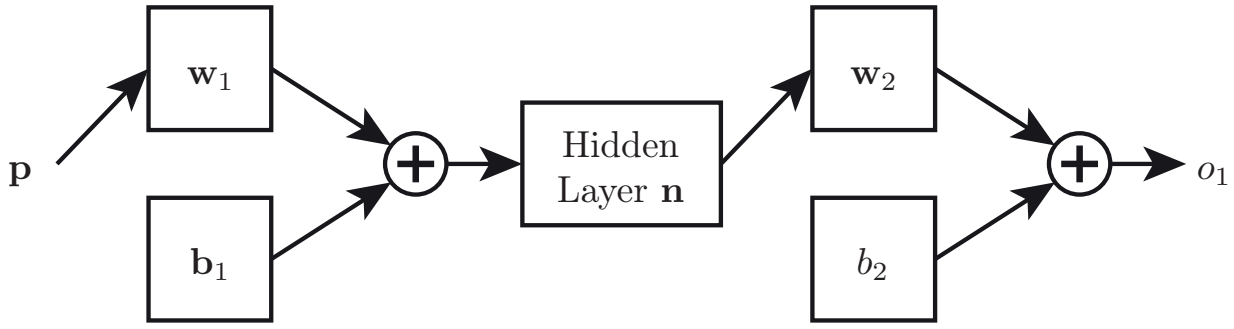


Figure C.1: Block diagram of the NN

The complete NN is given by

$$o_1 = \left(\frac{2}{1 + \mathrm{e}^{-2*(\mathbf{p}\cdot\mathbf{w}_1 + \mathbf{b}_1)}} - 1 \right) \cdot \mathbf{w}_2 + b_2 \,. \tag{C.1}$$

The weight matrix between the input parameters p_i and the ten neurons n_i of the first layer is given in Table C.1. The values represent the weights between the parameter given in the first row and the hidden neuron given in the first column (e.g., the first parameter p_1 is weighted by -1.7491 for the first hidden neuron n_1).

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
n_1	-1.7491	0.094	-0.8648	-1.0145	-0.6265	1.4537	-0.4723
n_2	0.6236	0.5349	-1.5471	0.1404	-0.5232	1.8527	-0.8162
n_3	-0.3902	-0.3654	-0.2579	0.0013	-0.2677	-2.3801	-1.6421
n_4	1.4473	0.0063	0.1861	-0.0115	-0.1129	-2.1217	0.61037
n_5	0.6962	-0.896	1.2199	-1.0235	-0.2289	-0.881	-0.8501
n_6	-3.0021	0.275	0.2443	0.847	1.8809	-0.5474	1.5235
n_7	-0.0745	1.2259	0.8441	1.3127	0.4133	-0.4619	-1.1074
n_8	0.6809	0.744	1.1811	-0.5306	0.1849	0.9648	-0.4005
n_9	-0.5373	-1.0064	-0.7305	-1.0939	0.2841	-0.7478	-0.3536
n_{10}	-0.8504	-1.2027	0.1598	-0.9955	-0.4099	-1.2165	-0.6226

Table C.1: Weight matrix between the input parameters and the neurons of the hidden layer

The bias vector $\mathbf{b}_1$ after the input layer is given in Table C.2.

	$\mathbf{b}_1$
n_1	1.5621
n_2	1.6424
n_3	1.4563
n_4	-0.3875
n_5	-0.0458
n_6	0.2504
n_7	0.2194
n_8	1.567
n_9	-2.1019
n_{10}	-2.4587

Table C.2: Bias vector $\mathbf{b}_1$ after the input layer

The weight vector $\mathbf{w}_2$ between the hidden layer and the output is given in Table C.3. Due to the length of the vector, it is transposed compared to the Tables C.1 and C.2 (e.g., the output of the hidden neuron n_1 is weighted by -0.5511 in the output).

	o_1
n_1	-0.5511
n_2	-1.4703
n_3	-1.6853
n_4	0.7285
n_5	-1.2577
n_6	-1.6339
n_7	-0.8382
n_8	-0.8472
n_9	-0.5111
n_{10}	-1.0219

Table C.3: Weight vector $\mathbf{w}_2$ between the hidden layer and the output neuron

The single bias b_2 after the hidden layer is given in Table C.4.

	b_2
o_1	0.65212

Table C.4: Bias after the hidden layer

Bibliography

[3GP88] 3GPP TS 06.10. "GSM Full Rate Speech Transcoding", 1988.

[3GP07] 3GPP. "Speech Codec Speech Processing Functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) Speech Codec; Conformance Testing". TS 26.274, 3rd Generation Partnership Project (3GPP), June 2007.

[AB79] J. B. Allen and D. A. Berkley. "Image method for efficiently simulating small-room acoustics". *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[ADTA01] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. "The CIPIC HRTF database". *Proc. IEEE Workshop the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.

[Arb07] Arberet, S. and Gribonval, R. and Bimbot, F. "A Robust Method to Count and Locate Audio Sources in a Stereophonic Linear Anechoic Mixture". *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 3, pp. III–745 –III–748, april 2007.

[BFS+13] C. Bulla, C. Feldmann, M. Schäfer, F. Heese, T. Schlien, and M. Schink. "High Quality Video Conferencing: Region of Interest Encoding and Joint Video/Audio Analysis". *International Journal On Advances in Telecommunications*, vol. 6, no. 3&4, pp. forthcoming, December 2013.

[BGN00] R. H. Byrd, J. C. Gilbert, and J. Nocedal. "A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming". *Mathematical Programming*, vol. 89, no. 1, pp. 149–185, 2000.

[BHCN06] J. Benesty, Y. Huang, J. Chen, and P. A. Naylor. "Adaptive Algorithms for the Identification of Sparse Impulse Responses". *Topics in Acoustic Echo and Noise Control*. 2006.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[Bis07] A. Biswas. *Advances in Perceptual Stereo Audio Coding Using Linear Prediction Techniques*. PhD thesis, Technische Universiteit Eindhoven, 2007.

[Bla97] J. Blauert. *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.

[Bla05] J. Blauert. *Communication Acoustics*. Springer, 2005.

[Bod95] M. Bodden. "Binaural modeling and auditory scene analysis". *Proceedings of WASPAA*, pp. 31–34, 1995.

[BS90] A. Böttcher and B. Silbermann. *Analysis of Toeplitz operators*. Springer Monographs in Mathematics. Springer, 1990.

[BS96] K. Brandenburg and G. Stoll. "ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio". *Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction*, 5 1996.

[Buc10] H. Buchner. *Broadband Adaptive MIMO Filtering: A Unified Treatment and Applications to Acoustic Human-machine Interfaces*. PhD thesis, Friedrich-Alexander Universität Erlangen-Nürnberg, 2010.

[BvdPKS05] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers. "Parametric Coding of Stereo Audio". *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 561917, 2005.

[BW01] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Digital Signal Processing. Springer, 2001.

[CD78] H. Colburn and N. I. Durlach. *Hearing*, chapter 11, pp. 467–518. Academic Press, New York, 1978.

[Che53] E. C. Cherry. "Some Experiments on the Recognition of Speech, with One and with Two Ears". *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[Com61] F. C. Commission. "Amendment of Part 3 of the Commission's Rules and Regulations to Permit FM Broadcast Stations to Transmit Stereophonic Programs on a Multiplex Basis", 1961.

[Cô11] N. Côté. *Integral and Diagnostic Intrusive Prediction of Speech Quality*. T-Labs Series in Telecommunication Services. Springer, 2011.

[DM03] S. Doclo and M. Moonen. "Design of Far-field and Near-field Broadband Beamformers using Eigenfilters". *Signal Processing*, vol. 83, no. 12, pp. 2641–2673, 2003.

[DT73] L. Dorren and J. Torczyner. "An Optimum Quadraphonic FM Broadcasting System". *IEEE Transactions on Broadcast and Television Receivers*, pp. 277–285, November 1973.

[Dur63] N. I. Durlach. "Equalization and Cancellation Theory of Binaural Masking-Level Differences". *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, 1963.

[ECG00] C. O. Etemoglu, V. Cuperman, and A. Gersho. "Speech Coding with an Analysis-by-Synthesis Sinusoidal Model". *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. III–1371 –III–1374, 2000.

[ETS09] ETSI, Rec. TS 26.171. "Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description", 2009.

[FR11] E. Fisher and B. Rafaely. "Near-Field Spherical Microphone Array Processing With Radial Filtering". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 256–265, February 2011.

[FZ07] H. Fastl and E. Zwicker. *Psychoacoustics: facts and models*. Springer series in information sciences. Springer, 2007.

[Gai93] W. Gaik. "Combined Evaluation of Interaural Time and Intensity Differences: Psychoacoustical Results and Computer Modeling". *J. Acoust. Soc. Am.*, vol. 94, pp. 98–110, 1993.

[GG04] R. L. Goldsworthy and J. E. Greenberg. "Analysis of Speech-Based Speech Transmission Index Methods with Implications for Nonlinear Operations". *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[GN98] R. Gray and D. Neuhoff. "Quantization". *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2325–2383, 1998.

[GSV11] B. Geiser, M. Schäfer, and P. Vary. "Binaural Wideband Telephony Using Steganography". *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, vol. 61 of *Studientexte zur Sprachkommunikation*, pp. 132–137. ITG, DEGA, TuDPress Verlag der Wissenschaften GmbH, September 2011.

[GV07] B. Geiser and P. Vary. "Backwards Compatible Wideband Telephony in Mobile Networks: CELP Watermarking and Bandwidth Extension". *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. IV, pp. 533–536, Honolulu, Hawai'i, USA, April 2007.

[GZR06] S. George, S. Zielinski, and F. Rumsey. "Feature Extraction for the Prediction of Multichannel Spatial Audio Fidelity". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1994–2005, 2006.

[HCD+11] C. Helmrich, P. Carlsson, S. Disch, B. Edler, J. Hilpert, M. Neusinger, H. Purnhagen, N. Rettelbach, J. Robilliard, and L. Villemoes. "Efficient Transform Coding of Two-Channel Audio Signals by Means of Complex-Valued Stereo Prediction". *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 497 –500, may 2011.

[Her04] J. Herre. "From Joint Stereo to Spatial Audio Coding - Recent Progress and Standardization". *Proceedings of the 7th Int. Conference on Digital Audio Effects*, 2004.

[HL10] S. Haykin and K. J. R. Liu. *Handbook on Array Processing and Sensor Networks*. Adaptive and Learning Systems for Signal Processing, Communications and Control Series. Wiley, 2010.

[HLS93] S. Haykin, J. Litva, and T. Shepherd. *Radar array processing*. Springer series in information sciences. Springer-Verlag, 1993.

[HM94] M. T. Hagan and M. B. Menhaj. "Training feedforward networks with the Marquardt algorithm". *Neural Networks, IEEE Transactions on*, vol. 5, no. 6, pp. 989–993, nov 1994.

[HNS88] J. Holdsworth and I. Nimmo-Smith. "Implementing a GammaTone Filter Bank". Technical report, Cambridge Electronic Design, MRC Applied Psychology Unit, 1988.

[HS71] T. Houtgast and H. J. M. Steeneken. "Evaluation of Speech Transmission Channels by Using Artificial Signals". *Acustica*, vol. 25, pp. 355–367, 1971.

[HSA+02] T. Houtgast, H. Steeneken, W. Ahnert, L. Braida, R. Drullman, J. Festen, K. Jacob, P. Mapp, S. McManus, K. Payton, R. Plomp, J. Verhave, and S. van Wijngaarden. *Past, Present and Future of the Speech Transmission Index*. TNO Human Factors, Soesterberg, The Netherlands, 2002.

[HSV+12] F. Heese, M. Schäfer, P. Vary, E. Hadad, S. M. Golan, and S. Gannot. "Comparison of Supervised and Semi-supervised Beamformers Using Real Audio Recordings". *Proceedings of IEEE 27-th Convention of Electrical and Electronics Engineers in Israel (IEEEI)*. IEEE, November 2012.

[HSWV13] F. Heese, M. Schäfer, J. Wernerus, and P. Vary. "Numerical Near Field Optimization of a Non-Uniform Sub-band Filter-and-Sum Beamformer". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, May 2013.

[Int03] International Electrotechnical Commission. "Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index". IEC 60268-16:2003, May 2003.

[ISO09] ISO/IEC. *Information technology – Coding of audio-visual objects – Part 3: Audio*. International Organization for Standardization, 2009.

[ITU90] ITU-T Rec. G.726. "40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)", 1990.

[ITU96a] ITU. *Methods for subjective determination of transmission quality (ITU-T Recommendation P.800)*. International Telecommunications Union, August 1996.

[ITU96b] ITU-T Rec. G.723.1. "G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s", 1996.

[ITU96c] ITU-T Rec. G.729. "G.729 : Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)", 1996.

[ITU01a] ITU. *Method for Objective Measurements of Perceived Audio Quality (ITU-R Recommendation BS.1387-1).* International Telecommunications Union, 2001.

[ITU01b] ITU. "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (ITU-T Recommendation P.862)", 2001.

[ITU03] ITU-T Rec. G.722.2. "Wideband Coding of Speech at Around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)", 2003.

[ITU05] ITU. "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs (ITU-T Recommendation P.862.2)", 2005.

[ITU06] ITU-T Rec. G.729.1. "G.729 Based Embedded Variable Bit-Rate Coder: An 8-32 kbit/s Scalable Wideband Coder Bitstream Interoperable with G.729.", 2006.

[ITU08] ITU-T Rec. G.718. "Frame Error Robust Narrowband and Wideband Embedded Variable Bit-Rate Coding of Speech and Audio from 8-32 kbit/s", 2008.

[ITU11] ITU. "Perceptual objective listening quality assessment (ITU-T Recommendation P.863)", 2011.

[Jef48] L. A. Jeffress. "A place theory of sound localization". *J. Comp. Physiol. Psych.*, vol. 41, pp. 35–39, 1948.

[JF92] J. Johnston and A. Ferreira. "Sum-difference stereo transform coding". *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-92*, vol. 2, pp. 569–572 vol.2, 1992.

[JHN+12] M. Jeub, C. Herglotz, C. M. Nelke, C. Beaugeant, and P. Vary. "Noise Reduction for Dual-Microphone Mobile Phones Exploiting Power Level Differences". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, March 2012. NEC best paper award.

[JN84] N. Jayant and P. Noll. *Digital Coding of Waveforms.* Prentice-Hall, Inc., 1984.

[JSEV10] M. Jeub, M. Schäfer, T. Esch, and P. Vary. "Model-Based Dereverberation Preserving Binaural Cues". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732 – 1745, September 2010. Special Issue on Processing Reverberant Speech.

[JSK+10] M. Jeub, M. Schäfer, H. Krüger, C. M. Nelke, C. Beaugeant, and P. Vary. "Do We Need Dereverberation for Hand-Held Telephony?". *International Congress on Acoustics (ICA)*. Australian Acoustical Society, August 2010.

[JSV09] M. Jeub, M. Schäfer, and P. Vary. "A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms". *Proc. 16th International Conference on Digital Signal Processing*, 2009.

[JWS07] B. R. John William Strutt. "On our Perception of Sound Direction". *Philosophical Magazine*, vol. 13, pp. 214–232, 1907.

[KAWW96] R. A. Kennedy, T. Abhayapala, D. B. Ward, and R. C. Williamson. "Nearfield Broadband Frequency Invariant Beamforming". *IEEE ICASSP*, 1996.

[KR86] P. Kabal and R. Ramachandran. "The computation of line spectral frequencies using Chebyshev polynomials". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 6, pp. 1419–1426, 1986.

[Kru10] H. Krueger. *Low Delay Audio Coding Based on Logarithmic Spherical Vector Quantization*. Dissertation, IND, RWTH Aachen, March 2010.

[Kut09] H. Kuttruff. *Room Acoustics*. Spon Press, Oxon, 5th edition, 2009.

[KV08] H. Krüger and P. Vary. "A New Approach for Low-Delay Joint-Stereo Coding". *ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, October 2008.

[LEKP90] C. Ludvigsen, C. Elberling, G. Keidser, and T. Poulsen. "Prediction of Intelligibility of Non-linearly Processed Speech". *Acta oto-laryngologica. Supplementum*, vol. 469, pp. 190–195, 1990.

[Lin86a] W. Lindemann. "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals". *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1608–1622, 1986.

[Lin86b] W. Lindemann. "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front". *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1623–1630, 1986.

[LVSH11] M. Lutzky, M. L. Valero, M. Schnell, and J. Hilpert. "AAC-ELD V2 - The New State of the Art in High Quality Communication Audio Coding". *Audio Engineering Society Convention 131*, 10 2011.

[MA76] J. Markel and A.Gray. *Linear Prediction of Speech*. Springer, 1976.

[Mak75] J. Makhoul. "Linear prediction: A tutorial review". *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561 – 580, april 1975.

[MHA08] R. Martin, U. Heute, and C. Antweiler, editors. *Advances in Digital Speech Transmission*. John Wiley & Sons, Ltd., January 2008.

[MHW01] R. Martin, C. Hoelper, and I. Wittke. "Estimation of Missing LSF Parameters Using Gaussian Mixture Models". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 729–732, May 2001.

[MKK94] N. Murphy, A. Krukowski, and I. Kale. "Implementation of wideband integer and fractional delay element". *Electronics Letters*, vol. 30, no. 20, pp. 1658 –1659, sep 1994.

[Moo20] E. H. Moore. "On the reciprocal of the general algebraic matrix". 1920.

[MW72] J. I. Makhoul and J. J. Wolf. "Linear prediction and the spectral analysis of speech". BBN Report 2304, Bolt Beranek and Newman Inc., Boston, Massachusetts, August 1972.

[MZ93] S. Mallat and Z. Zhang. "Matching pursuits with time-frequency dictionaries". *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397 –3415, dec 1993.

[NA94] NTT-AT. "Multi-lingual speech database for telephonometry", 1994.

[NLT89] A. K. Nábělek, T. R. Letowski, and F. M. Tucker. "Reverberant overlap- and self-masking in consonant identification". *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259–1265, October 1989.

[Pol88] J.-D. Polack. *La transmission de l'energie sonore dans les salles*. PhD thesis, Universite du Maine, Le Mans, France, 1988.

[PT55] R. Penrose and J. A. Todd. "A generalized inverse for matrices". *Mathematical Proceedings of The Cambridge Philosophical Society*, vol. 51, 1955.

[Raa80] J. Raatgever. *On the Binaural Processing of Stimuli with Different Interaural Phase Relations*. PhD thesis, Technische Hogeschool Delft, The Netherlands, 1980.

[RBK+06] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza. "Objective Assessment of Speech and Audio Quality – Technology and Applications". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1890–1901, 2006.

[RG97] J. G. Ryan and R. A. Goubran. "Near-Field Beamforming for Microphone Arrays". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[RG00] J. G. Ryan and R. A. Goubran. "Array Optimization Applied in the Near Field of a Microphone Array". *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 173–176, March 2000.

[RKV12] C. Rohlfing, H. Krüger, and P. Vary. "logarithmic cubic vector quantization: Concept and analysis". *Proceedings of International Symposium on Information Theory and its Applications (ISITA)*, pp. 294–298. IEICE, October 2012.

[RV87] D. D. Rife and J. Vanderkooy. "Transfer-Function Measurement Using Maximum-Length Sequences". *Audio Engineering Society Convention 83*, 10 1987.

[Sau13] B. Sauert. *Near-End Listening Enhancement: Theory and Application*. PhD thesis, RWTH Aachen University, 2013.

[SBV12] M. Schäfer, M. Bahram, and P. Vary. "Improved Binaural Model for Localization of Multiple Sources". *Proceedings of 10. ITG Symposium Speech Communication*, 2012.

[SBV13] M. Schäfer, M. Bahram, and P. Vary. "An Extension of the PEAQ Measure by a Binaural Hearing Model". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, May 2013.

[Sch12] D. Schröder. *Physically based real-time auralization of interactive virtual environments*. PhD thesis, RWTH Aachen University, 2012.

[SEV06] B. Sauert, G. Enzner, and P. Vary. "Near End Listening Enhancement with Strict Loudspeaker Output Power Constraining". *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*. IWAENC, September 2006.

[SH80] H. J. M. Steeneken and T. Houtgast. "A physical method for measuring speech-transmission quality". *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, January 1980.

[SHS+13] T. Schlien, F. Heese, M. Schäfer, C. Antweiler, and P. Vary. "Audiosignalverarbeitung für Videokonferenzsysteme". *Workshop Audiosignal- und Sprachverarbeitung (WASP)*. Gesellschaft für Informatik, September 2013. Workshop im Rahmen der 43. Jahrestagung der Gesellschaft für Informatik.

[SHWV12] M. Schäfer, F. Heese, J. Wernerus, and P. Vary. "Numerical Near Field Optimization of Weighted Delay-and-Sum Microphone Arrays". *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*. IWAENC, September 2012.

[SJ84] F. Soong and B. Juang. "Line spectrum pair (LSP) and speech data compression". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9, pp. 37–40, 1984.

[SJSV10] M. Schäfer, M. Jeub, B. Sauert, and P. Vary. "Reverberation-Based Post-Processing for Improving Speech Intelligibility". *International Congress on Acoustics (ICA)*. Australian Acoustical Society, August 2010.

[SKV09] M. Schäfer, H. Krüger, and P. Vary. "Extending Monaural Speech and Audio Codecs by Inter-Channel Linear Prediction". *20. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, vol. 1, 2009.

[SMH92] T. M. Shackleton, R. Meddis, and M. J. Hewitt. "Across frequency integration in a model of lateralization". *J. Acoust. Soc. Am.*, vol. 91, no. 4, pp. 2276–2279, 1992.

[SSGV11] M. Schäfer, C. Schnelling, B. Geiser, and P. Vary. "A Listening Test Environment for Subjective Assessment of Speech and Audio Signal Processing Algorithms". *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, vol. 61 of *Studientexte zur Sprachkommunikation*, pp. 237–244. ITG, DEGA, TUDpress Verlag der Wissenschaften GmbH, September 2011.

[SV10] R. Scharrer and M. Vorländer. "Blind Reverberation Time Estimation". *Proceedings ICA 2010, 20th International Congress on Acoustics : 23 - 27 August 2010, Sydney, New South Wales, Australia / [Eds.: Marion Burgess ...]*. Australian Acoustical Society, NSW Division, 2010. 1 CD-ROM.

[SV12a] B. Sauert and P. Vary. "Near-End Listening Enhancement in the Presence of Bandpass Noises". *ITG-Fachtagung Sprachkommunikation*, pp. 195–198. VDE Verlag GmbH, September 2012.

[SV12b] M. Schäfer and P. Vary. "Hierarchical Multi-Channel Audio Coding based on Time-Domain Linear Prediction". *Proceedings of European Signal Processing Conference (EUSIPCO)*, pp. 2148–2152. EURASIP, August 2012.

[SZT88] R. M. Stern, A. S. Zeiberg, and C. Trahiotis. "Lateralization of complex binaural stimuli: A weighted-image model". *J. Acoust. Soc. Am.*, vol. 84, no. 1, pp. 156–165, 1988.

[TF06] B. G. Tabachnick and L. S. Fidell. *Using Multivariate Statistics (5th Edition)*. Allyn & Bacon, 5 edition, March 2006.

[VL93] V. Valimaki and T. Laakso. "Fractional delay digital filters". *Circuits and Systems, 1993., ISCAS '93, 1993 IEEE International Symposium on*, pp. 355 –359 vol.1, may 1993.

[VM06] P. Vary and R. Martin. *Digital Speech Transmission – Enhancement, Coding and Error Concealment*. John Wiley & Sons, Inc., Chichester, UK, 2006. ISBN 0-471-56018-9.

[VMR+88] T. Vogl, J. Mangis, A. Rigler, W. Zink, and D. Alkon. "Accelerating the convergence of the back-propagation method". *Biological Cybernetics*, vol. 59, pp. 257–263, 1988. 10.1007/BF00332914.

[VTMM10] J.-M. Valin, T. B. Terriberry, C. Montgomery, and G. Maxwell. "A High-Quality Speech and Audio Codec With Less Than 10-ms Delay". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 58–67, 2010.

[Wel67] P. Welch. "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified

periodograms". *Audio and Electroacoustics, IEEE Transactions on*, vol. 15, no. 2, pp. 70 – 73, jun 1967.

[WGH+06] J. Wen, N. Gaubitch, E. Habets, T. Myatt, and P. Naylor. "Evaluation of Speech Dereverberation Algorithms using the MARDY Database". *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, 2006.

[YC90] T. C. Yin and J. C. Chan. "Interaural time sensitivity in medial superior olive of cat". *J. Neurophysiol.*, vol. 64, pp. 465–488, 1990.

[Yul27] G. U. Yule. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 226, pp. 267–298, april 1927.

[ZF67] E. Zwicker and R. Feldtkeller. *Das Ohr als Nachrichtenempfänger*. S. Hirzel Verlag, 1967.

[ZGET04] Y. Zheng, R. Goubran, and M. El-Tanany. "Robust near-field adaptive beamforming with distance discrimination". *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 478 – 488, sept. 2004.

[Zha10] F. Zhang. *The Schur Complement and Its Applications*. Numerical Methods and Algorithms. Springer, 2010.

[ZRKB05] S. Zielinski, F. Rumsey, R. Kassier, and S. Bech. "Development and initial validation of a multichannel audio quality expert system". *Journal of the Audio Engineering Society*, vol. 53, no. 1/2, pp. 4–21, 2005.