Graz University of Technology

Doctoral Thesis

Speech Watermarking and Air Traffic Control

Konrad Hofbauer

Faculty of Electrical and Information Engineering
Graz University of Technology, Austria

Advisor: Prof. DI Dr. Gernot Kubin (TU Graz)
Co-Advisor: Prof. DDr. W. Bastiaan Kleijn (KTH Stockholm)

Graz, March 6, 2009

Abstract

Air traffic control (ATC) voice radio communication between aircraft pilots and controllers is subject to technical and functional constraints owing to the legacy radio system currently in use worldwide. This thesis investigates the embedding of digital side information, so called watermarks, into speech signals. Applied to the ATC voice radio, a watermarking system could overcome existing limitations, and ultimately increase safety, security and efficiency in ATC. In contrast to conventional watermarking methods, this field of application allows embedding of the data in perceptually irrelevant signal components. We show that the resulting theoretical watermark capacity far exceeds the capacity of conventional watermarking channels. Based on this finding, we present a general purpose blind speech watermarking algorithm that embeds watermark data in the phase of non-voiced speech segments by replacing the excitation signal of an autoregressive signal representation. Our implementation embeds the watermark in a subband of narrowband speech at a watermark bit rate of the order of magnitude of 500 bit/s. The system is evaluated experimentally using an ATC speech corpus and radio channel measurement results, both of which were produced specifically for this purpose and constitute contributions on their own. The adaptive equalization based watermark detector is able to recover the watermark data in the presence of channels with non-linear phase, time-variant bandpass filtering, amplitude gain modulation, desynchronization and additive noise. In the aeronautical application the scheme is highly robust and outperforms current state-of-the-art speech watermarking methods. The theoretical results show that the performance could be increased even further by incorporating perceptual masking.

Keywords: Information hiding, digital watermarking, speech watermarking, legacy system enhancement, watermark capacity, watermark synchronization, air traffic control, voice radio, analog radio channel.

Zusammenfassung

Die Sprechfunk-Kommunikation zwischen Piloten und Fluglotsen unterliegt technischen und funktionalen Einschränkungen, die sich aus dem althergebrachten und weltweit eingesetzten analogen Sprechfunk-System ergeben. Diese Arbeit beschäftigt sich mit der Einbettung digitaler Seiteninformation, so genannter Wasserzeichen, in Sprachsignale. Angewandt auf den analogen Flugfunk könnte ein solches Wasserzeichen-System bestehende Einschränkungen aufheben und letztendlich die Effizienz sowie die aktive und passive Sicherheit in der Flugsicherung steigern. Im Gegensatz zu herkömmlichen Wasserzeichen-Verfahren erlaubt dieser Einsatzbereich ein Einbetten der Daten in Signalanteilen, die vom menschlichen Gehör nicht wahrgenommen werden können. Es zeigt sich, dass dies zu einer erheblichen Steigerung der Kapazität des verdeckten Datenkanals führt. Basierend auf dieser Erkenntnis wird ein universelles blindes Wasserzeichen-Verfahren vorgestellt, das die Daten in die Phase von stimmlosen Sprachlauten einbettet, indem das Anregungssignal eines autoregressiven Signalmodells durch ein Wasserzeichen-Signal ersetzt wird. Die Implementierung bettet die Seiteninformationen mit einer Bitrate von circa 500 bit/s in einem Teilband des schmalbandigen Sprachsignals ein. Zur experimentellen Validierung des Systems wurden Flugfunkkanal-Messungen durchgeführt und ein für die Flugsicherung spezifischer Sprachkorpus erstellt. Diese beiden Datensammlungen stellen jeweils einen eigenständigen Beitrag dar. Der Wasserzeichen-Detektor basiert auf adaptiver Entzerrung und detektiert die eingebetteten Daten sogar bei einer Übertragung des Signals über Kanäle mit zeitvarianter Kanalverstärkung, nicht-linearer Phase, zeitvarianter Bandpass-Filterung, Desynchronisierung und additivem Rauschen. Das Verfahren ist hochrobust und übertrifft in der Flugfunk-Anwendung leistungsmäßig die modernsten existierenden Wasserzeichen-Verfahren. Die vorliegenden theoretischen Ergebnisse zeigen, dass durch eine Berücksichtigung von gehörbezogenen Maskierungsmodellen die Leistungsfähigkeit weiter gesteigert werden kann.

Stichwörter: Digitale Wasserzeichen, Einbettung digitaler Seiteninformationen, Sprachwasserzeichen, Wasserzeichenkanalkapazität, Wasserzeichensynchronisierung, Flugsicherung, Flugfunk, Sprechfunk, analoger Sprechfunkkanal

Acknowledgments

Many people have contributed to my journey towards this dissertation. It is with great pleasure that I take this opportunity to acknowledge the support I have received.

I am most deeply indebted to my mentor Horst Hering at Eurocontrol. The basic idea of this thesis, the application of watermarking to ATC, is a brainchild of yours, and without your initiative and persistence this entire project would have never been possible. Your keen interest and belief in my work and your continuing support and encouragement has always meant a lot to me. Many thanks also to Prof. Vu Duong, Marc Bourgois and Martina Jürgens of the former INO team at Eurocontrol for the continuing support.

I am equally indebted to my advisors Prof. Gernot Kubin and Prof. Bastiaan Kleijn. Your knowledge, creativity and keen intuitive insights, your dedication, professionalism and hard-working nature, as much as your hospitality, flexibility, patience and placidness have truly impressed and inspired me. Your guidance not only had a major impact on the outcome of this work, but also contributed greatly to my intellectual and personal growth.

The journey towards my degree had also been a journey throughout Europe, and I had the chance to work at and experience three different laboratories. A deserved thanks is given to my colleagues of the SPSC lab at TU Graz, the SIP group at KTH Stockholm, and the former INO team at Eurocontrol, for the meaningful discussions and for creating an enjoyable working environment. The names are too many to mention, but I am happy to give my warm thanks to all my former office mates, namely Tuan, Sarwar, Erhard and many more at SPSC, Martin, Daniel, Mario and Eri at INO, and Ermin and Gustav at SIP.

My deepest thanks to my parents Konrad and Annemarie who always encouraged and supported me throughout the many years of my studies. Finally, my warmest thanks to my beloved girlfriend Agata for your incredible love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Motivation

The advances in modern information technology revolutionized the way we communicate. Broadband Internet access and mobile wireless communication, be it for voice, data, or video transmission, are ubiquitous and seemingly available anytime and anywhere. Unfortunately, this is not entirely correct. The world is full of legacies, and it is not always possible to deploy or use state-of-the-art communication technologies.

The problem addressed in this thesis is the voice radio communication between aircraft and civil air traffic control (ATC) operators. Still today, this communication is based on pre-World War II technology using amplitude-modulation analog radio transmissions on shared broadcast channels. While technology has rapidly advanced and satellite communication facilitates broadband Internet access in the passenger cabin, the air/ground voice radio for pilots and controller communication has not been changed since its introduction in the early forties of the last century.

The large technological gap between communication systems available in the cabin for passenger services and in the cockpit for air traffic control purposes is not likely to disappear in the foreseeable future. This is due to a wide range of reasons, including an enormous base of airborne and ground legacy systems worldwide, long life-cycles for aeronautical systems, difficult technological requirements (such as large system scales and safety-critical compatibility, availability and robustness requirements), international regulations, and, last but not least, a lack of common political willingness or determination.

The ATC voice radio lacks basic features that are taken for granted in any modern communication system, such as caller identification or selective calling. It is commonly agreed that such features could improve safety, security and efficiency in air traffic control. Since, despite its advantages, there is no foreseeable time frame for the

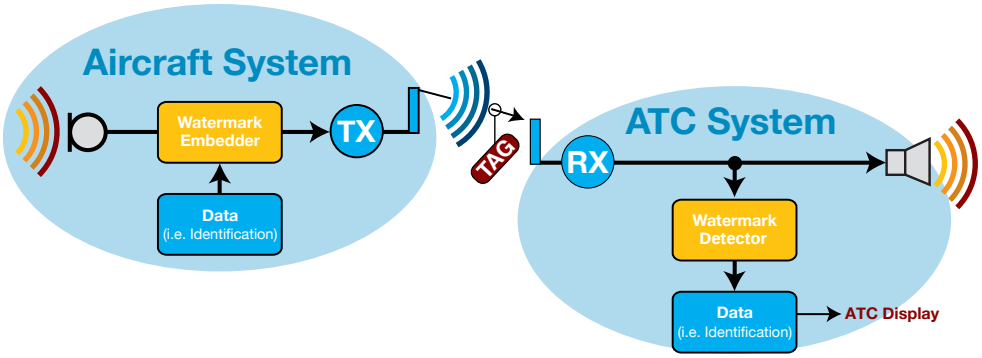


Figure 1.1.: Speech watermarking system for the air/ground voice radio.

introduction of digital ATC voice radio communication, it is attractive to retrofit or extend the current analog system with additional features.

Many potential features, such as caller identification, authentication, selective calling, call routing, or position reporting, can conceptually be reduced to a transmission of digital side information that accompanies the transmitted analog speech signal. In order to be legacy system compatible and to avoid the necessity of a separate transmitter, such side information should be transmitted in-band within the speech channel as shown in Figure 1.1. On the receiving end, such side information should not be noticeable also with unequipped legacy receivers.

The imperceptible embedding of side information into another signal, often referred to as information (or data) hiding or digital watermarking, constitutes the core of this thesis. We deal with the general problem of imperceptibly embedding digital data into speech, and consider the special case of embedding into ATC speech that is transmitted over an analog aeronautical radio channel.

Watermarking speech signals with a high embedded data rate is a difficult problem due to the narrow bandwidth of speech. The transmission of the watermarked speech over a poor-quality aeronautical radio channel poses an additional difficulty due to narrow bandwidth, noise corruption, fading and time-variance of the channel.

Compared to image, video and audio watermarking, there is relatively little prior work in the field of speech watermarking. However, the production and perception of speech is well understood, and one can draw from a large pool of knowledge gained in the context of speech coding.

This thesis covers a wide range of aspects of the aforementioned problem. It considers the full communication chain from source to sink, including the particular characteristics of the input speech, the embedding of the watermark, the characteristics of the aeronautical radio channel, and the detection of the watermark data in the degraded signal. Solutions are provided for a wide range of sub-problems.

The first part of this thesis deals with speech watermarking in a general context. It investigates the theoretical watermark capacity in speech signals and proposes a new speech watermarking method and a complete system implementation, which outperform the current state-of-the-art methods.

The second part presents contributions to the domain of air traffic control. In particular, it presents an empirical study of the aeronautical voice radio channel, a database of ATC operator speech, and an evaluation of the robustness of the proposed watermarking system in the aeronautical application.

1.2. Thesis Outline

This thesis is divided into two parts.

Part I considers the general problem of embedding digital side information in speech signals. After a short review of related work in **Chapter 2**, the following three chapters address the issues of watermark capacity estimation, robust and high-rate watermarking, and watermark synchronization.

Chapter 3 investigates the theoretical watermark capacity in speech given an additive white Gaussian noise transmission channel. Starting from the general capacity of the ideal Costa scheme, it shows that a large improvement in theoretical watermark capacity is possible if the application at hand allows watermarking in perceptually irrelevant signal components. Different ways to derive approximate and exact expressions for the watermark capacity when modulating the signal's DFT phase are presented and compared to an estimation of the capacity when modulating host signal frequency components that are perceptually masked. Parameters required for the experimental comparison of both methods are estimated from a database of ATC speech signals, which is presented in Chapter 7.

Chapter 4 presents a blind speech watermarking algorithm that embeds the watermark data in the phase of non-voiced speech by replacing the excitation signal of an autoregressive speech signal representation. The watermark signal is embedded in a frequency subband, which facilitates robustness against bandpass filtering channels. We derive several sets of pulse shapes that prevent intersymbol interference and that allow the creation of the passband watermark signal by simple filtering. A marker-based synchronization scheme robustly detects the location of the embedded watermark data without the occurrence of insertions or deletions.

In light of the potential application to analog aeronautical voice radio communication, we present experimental results for embedding a watermark in narrowband speech at a bit rate of 450 bit/s. The adaptive equalization-based watermark detector not only compensates for the vocal tract filtering, but also recovers the watermark data in the presence of non-linear phase and bandpass filtering, amplitude modulation and additive noise, making the watermarking scheme highly robust.

Chapter 5 discusses different aspects of the synchronization between watermark embedder and detector. We examine the issues of timing recovery and bit synchronization, the synchronization between the synthesis and the analysis systems, as well as the

data frame synchronization. Bit synchronization and synthesis/analysis synchronization are not an issue when using the adaptive equalization-based watermark detector of Chapter 4. For the simpler linear prediction-based detector we present a timing recovery mechanism based on the spectral line method which achieves near-optimal performance.

Using a fixed frame grid and the embedding of preambles, the information-carrying frames are detected in the presence of preamble bit errors with a ratio of up to 10 %. Evaluated with the full watermarking system, the active frame detection performs near-optimal with the overall bit error ratio increasing by less than 0.5 %-points compared to ideal synchronization.

Part II of this thesis contains contributions to the domain of air traffic control (ATC). After a brief overview of the application of speech watermarking in ATC and related work in **Chapter 6**, the subsequent chapters present a radio channel measurement system, database and model, an ATC speech corpus, and an evaluation of the robustness of the proposed watermarking application.

Chapter 7 presents the ATCOSIM Air Traffic Control Simulation Speech corpus, a speech database of ATC operator speech. ATCOSIM is a contribution to ATC-related speech corpora. It consists of ten hours of speech data, which were recorded during ATC real-time simulations. The database includes orthographic transcriptions and additional information on speakers and recording sessions. The corpus is publicly available and provided free of charge. Possible applications of the corpus are, among others, ATC language studies, speech recognition and speaker identification, the design of ATC speech transmission systems, as well as listening tests within the ATC domain.

Chapter 8 presents a system for measuring time-variant channel impulse responses and a database of such measurements for the aeronautical voice radio channel. Maximum length sequences (MLS) are transmitted over the voice channel with a standard aeronautical radio and the received signals are recorded. For the purpose of synchronization, the transmitted and received signals are recorded in parallel to GPS-based timing signals. The flight path of the aircraft is accurately tracked. A collection of recordings of MLS transmissions is generated during flights with a general aviation aircraft. The measurements cover a wide range of typical flight situations as well as static back-to-back calibrations. The resulting database is available under a public license free of charge.

Chapter 9 proposes a data model to describe the data in the *TUG-EEC-Channels* database, and a corresponding estimation method. The model is derived from various effects that can be observed in the database, such as different filter responses, a time-variant gain, a sampling frequency offset, a DC offset and additive noise. To estimate the model parameters, we compare six well-established FIR filter identification techniques and conclude that best results are obtained using the method of Least

Squares. We also provide simple methods to estimate and compensate the sampling frequency offset and the time-variant gain.

The data model achieves a fit with the measured data down to an error of -40 dB, with the modeling error being smaller than the channel's noise. Applying the model to select parts of the database, we conclude that the measured channel is frequency-nonselective. The data contains a small amount of gain modulation (flat fading). Its source could not be conclusively established, but several factors indicate that it is *not* a result of radio channel fading. The observed noise levels are in a range from 40 dB to 23 dB in terms of SNR.

Chapter 10 shows the robustness of the proposed watermarking method in the aeronautical application using the channel model derived in Chapter 9. We experimentally demonstrate the robustness of the method against filtering, desynchronization, gain modulation and additive noise. Furthermore we show that pre-processing of the speech signal with a dynamic range controller can improve the watermark robustness as well as the intelligibility of the received speech.

Chapter 11 concludes both parts of this thesis and suggests directions for future research.

1.3. Contributions

The main contributions of this thesis are:

- a novel speech watermarking scheme that outperforms current state-of-the-art methods and is robust against AWGN, gain modulation, time-variant filtering and desynchronization
  - presented in Chapter 4 and Chapter 10
  - published in [1][2][3]
- methods for watermark synchronization on the sampling, bit and frame level
  - presented in Chapter 5
  - published in [1][2]
- watermark capacity estimations for phase modulation based and frequency masking based speech watermarking
  - presented in Chapter 3
  - recent results, unpublished
- a battery-powered portable measurement system for mobile audio channels
  - presented in Chapter 8
  - published in [4]
- a database of aeronautical voice radio channel measurements

  - presented in Chapter 8
  - published in [4]
  - publicly available at `http://www.spsc.tugraz.at/TUG-EEC-Channels`
- a data model and an estimation technique for measured aeronautical voice radio channel data
  - presented in Chapter 9
  - recent results, unpublished
- an air traffic control simulation speech corpus
  - presented in Chapter 7
  - published in [5]
  - publicly available at `http://www.spsc.tugraz.at/ATCOSIM`

1.4. List of Publications

First-Author Publications

For the following papers, the author of this thesis did the major part of the theoretical work, conducted all experiments, and did the major part in writing the paper.

[1] K. Hofbauer, G. Kubin, and W. B. Kleijn, "Speech watermarking for analog flat-fading bandpass channels," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, revised and resubmitted.

[2] K. Hofbauer and H. Hering, "Noise robust speech watermarking with bit synchronisation for the aeronautical radio," in *Information Hiding*, ser. Lecture Notes in Computer Science. Springer, 2007, vol. 4567/2008, pp. 252–266.

[3] K. Hofbauer and G. Kubin, "High-rate data embedding in unvoiced speech," in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, Pittsburgh, PY, USA, Sep. 2006, pp. 241–244.

[4] K. Hofbauer, H. Hering, and G. Kubin, "A measurement system and the TUG-EEC-Channels database for the aeronautical voice radio," in *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Montreal, Canada, Sep. 2006, pp. 1–5.

[5] K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM corpus of non-prompted clean air traffic control speech," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008.

[6] K. Hofbauer and G. Kubin, "Aeronautical voice radio channel modelling and simulation—a tutorial review," in *Proceedings of the International Conference on Research in Air Transportation (ICRAT)*, Belgrade, Serbia, Jul. 2006.

[7] K. Hofbauer, H. Hering, and G. Kubin, "Speech watermarking for the VHF radio channel," in *Proceedings of the EUROCONTROL Innovative Research Workshop (INO)*, Brétigny-sur-Orge, France, Dec. 2005, pp. 215–220.

[8] K. Hofbauer and H. Hering, "Digital signatures for the analogue radio," in *Proceedings of the NASA Integrated Communications Navigation and Surveillance Conference (ICNS)*, Fairfax, VA, USA, 2005.

[9] K. Hofbauer, "Advanced speech watermarking for secure aircraft identification," in *Proceedings of the EUROCONTROL Innovative Research Workshop (INO)*, Brétigny-sur-Orge, France, Dec. 2004.

Co-Author Publications

For the following paper, the author of this thesis helped with and guided the theoretical work, the experiments, and the writing of the paper.

[10] M. Gruber and K. Hofbauer, "A comparison of estimation methods for the VHF voice radio channel," in *Proceedings of the CEAS European Air and Space Conference (Deutscher Luft- und Raumfahrtkongress)*, Berlin, Germany, Sep. 2007.

The work of the following two papers is not included in this thesis. The author helped with the implementation of the experiments and the writing of the papers.

[11] H. Hering and K. Hofbauer, "From analogue broadcast radio towards end-to-end communication," in *Proceedings of the AIAA Aviation Technology, Integration, and Operations Conference (ATIO)*, Anchorage, Alaska, USA, Sep. 2008.

[12] H. Hering and K. Hofbauer, "Towards selective addressing of aircraft with voice radio watermarks," in *Proceedings of the AIAA Aviation Technology, Integration, and Operations Conference (ATIO)*, Belfast, Northern Ireland, Sep. 2007.

Part I.

Speech Watermarking

Chapter 2

Background and Related Work

2.1. Digital Watermarking

Watermarking is the process of altering a work or data stream to embed an additional message that is not or hardly perceptible. In digital watermarking, the work or data stream to be altered is called the 'host signal', and may be an image, video, audio, text or speech signal. A generic model for digital watermarking is shown in Figure 2.1. The 'watermark' is the embedded message and is expected to cause minimal perceptual degradation to the host signal.[1] Any modification to the watermarked signal in between the watermark embedding and the watermark detection constitutes a 'channel attack' and might render the watermark undetectable. The watermark capacity is the maximum information rate of the embedded message given the host signal, the allowed perceptual distortion induced by the watermark, and the watermark detection errors induced by the channel attack.

A brief and very accessible introduction to watermarking including a presentation of early schemes can be found in [14]. An exhaustive treatment of the underlying concepts of watermarking is provided in [13].

Practical applications for digital watermarking range from copy prevention and traitor tracing to broadcast monitoring, archiving, and legacy system enhancement. We only consider the case of blind watermarking, where the original host signal is not available at the watermark detector. Most watermarking methods are subject to a fundamental trade-off between watermark capacity, perceptual fidelity and robustness against intentional or unintentional channel attacks. The selection of an appropriate operating point is highly application-dependent.

[1] In [13], the term 'watermarking' is used only if the embedded message is related to the host signal. Otherwise, the term 'overt embedded communications' is used if the existence of an embedded message is known, or 'steganography' if the existence of the embedded message is concealed.

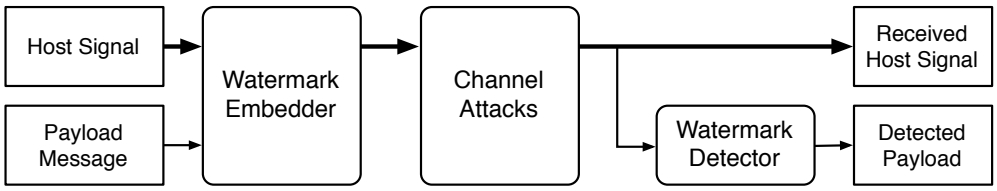


Figure 2.1.: Generic watermarking model.

2.2. Related Work—General Watermarking

In recent years, digital watermarking techniques achieved significant progress [13, 15]. Early methods considered watermarking as a purely additive process (Figure 2.2a), with potential spectral shaping of the watermark signal (Figure 2.2b), or an additive embedding of the watermark in a transform domain of the host signal (Figure 2.2c). As an example, spread spectrum watermarking adds a pseudo-random signal vector representing a watermark message to a transformation of the original host signal. This type of system was used in many practical implementations and proved to be highly robust, but suffered from the inherent interference between the watermark and the host signal [16, 17, 18, 19].

Costa's work on communication with side information [20] enabled a breach with traditional additive watermarking. It allows to consider watermarking as a communication problem, with the host signal being side information that is available to the embedder but not to the detector. The *watermark* channel capacity then becomes

$$C_{\mathrm{W,\,Costa}} = \frac{1}{2}\log_2\left(1+\frac{\sigma_{\mathrm{W}}^2}{\sigma_{\mathrm{N}}^2}\right) \tag{2.1}$$

and depends solely on the watermark-to-attack-noise power ratio $\frac{\sigma_{\mathrm{W}}^2}{\sigma_{\mathrm{N}}^2}$, assuming independent identically distributed (IID) host data, a power-constrained watermark signal and an additive white Gaussian noise (AWGN) channel attack [20]. The capacity is independent of the host signal power σ_{H}^2, and thus watermarking without host signal interference is possible [21, 22]. Striving to approach this capacity with practical schemes, quantization-based methods embed the information by requantizing the signal (Figure 2.2d) or a transform domain representation (Figure 2.2e) using different vector quantizers or dither signals that depend on the watermark data. These methods are often referred to as quantization index modulation (QIM). Reducing the embedding dimensionality of distortion-compensated QIM to one (sample-wise embedding) and using scalar uniform quantizers in combination with a watermark scale factor results in the popular scalar Costa scheme (SCS) [23]. SCS is optimally robust against the AWGN attack, but is vulnerable to amplitude scaling. A number of methods, for example the rational dither modulation scheme [24], have been proposed to counteract this vulnerability. However, quantization-based methods are still sensitive to many simple signal processing operations such as linear filtering, non-linear operations, mixing or resampling, which are the subjects of numerous recent investigations (e.g. [25], [26]).

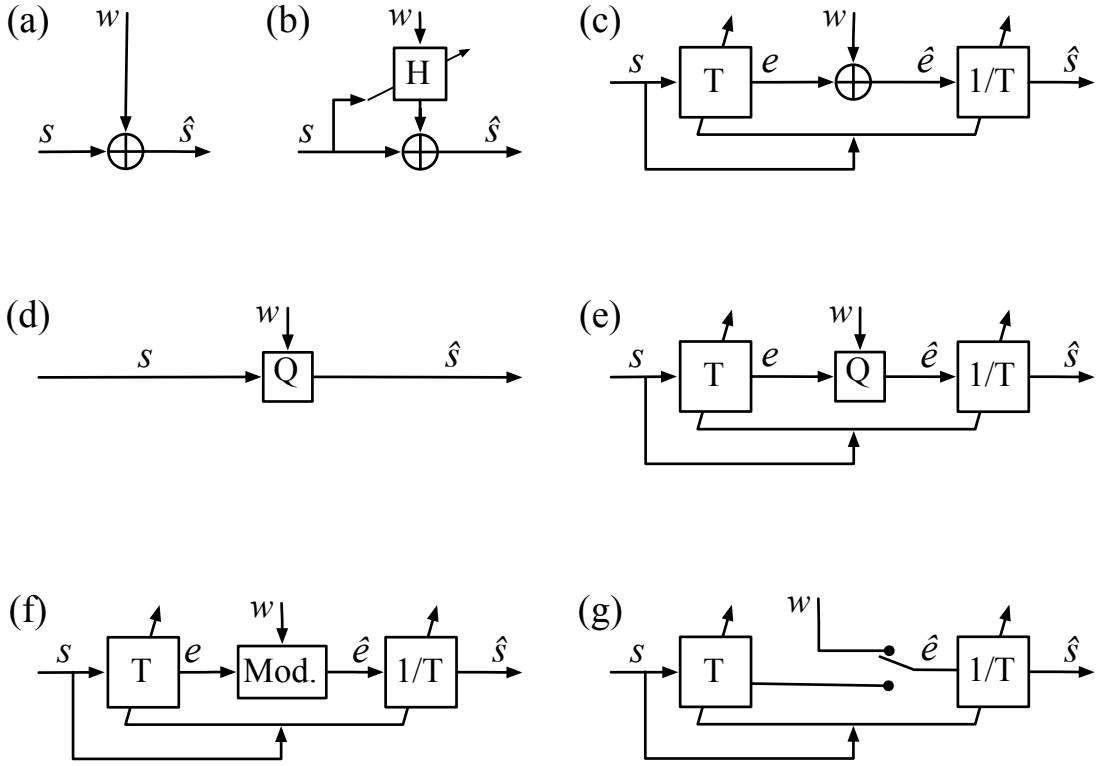


Figure 2.2.: A watermarked signal $\hat{s}(n)$ is generated from the original speech signal $s(n)$ and the watermark data signal $w(n)$ by **(a)** adding the data to $s(n)$, **(b)** adding adaptively filtered data to $s(n)$, **(c)** adding the data to a transform domain signal representation $e(n)$, **(d)** quantizing $s(n)$ according to the data, **(e)** quantizing the transform $e(n)$ according to the data, **(f)** modulating $e(n)$ with the data, and, as proposed in this work, **(g)** replacing parts of $e(n)$ by the data.

2.3. Related Work—Speech Watermarking

Most watermarking methods use a perceptual model to determine the permissible amount of embedding-induced distortion. Many audio watermarking algorithms use auditory masking, and most often frequency masking, as the perceptual model for watermark embedding. While this is a valid approach also for speech signals, we claim that capacity remains unused by not considering the specific properties of speech and the way it is perceived. It is thus indispensable to tailor a speech watermarking algorithm to the specific signal.

In speech watermarking, as in audio watermarking, spread-spectrum type systems used to be the method of choice for robust embedding [27, 28]. Various quantization-based methods have been proposed in recent years, applying QIM or SCS to autoregressive model parameters [29, 30], the pitch period [31], discrete Hartley transform coefficients [32], and the linear prediction residual [33]. The achievable bit rates range from a few bits to a few hundred bits per second, with varying robustness against different types of attacks. In a generalization of QIM, some methods modulate the speech signal or one of its components according to the watermark data (Figure 2.2f) by estimating and adjusting the polarity of speech syllables [34], by modulating the frequency of selected partials of a sinusoidal speech or audio model [35, 36], or by modifying the linear prediction model coefficients of the speech signal using long analysis and synthesis windows [37]. Taking quantization and modulation one step further, one can even replace certain speech components by a perceptually similar watermark signal (Figure 2.2g). Methods have been proposed that exchange time-frequency components of audio signals above 5 kHz that have strong noise-like properties by a spread spectrum sequence [38], or replace signal components that are below a perceptual masking threshold by a watermark signal [39, 40].

The above-mentioned algorithms are limited either in terms of their embedding capacity or in their robustness against the transmission channel attacks expected in the aeronautical application. This is due to the fact that the methods are designed for particular channel attacks (such as perceptual speech coding), focus on security considerations concerning hostile channel attacks, or fail to thoroughly exploit state-of-the-art watermarking theory or the characteristic features of speech perception.

2.4. Our Approach

In our approach, we combine the watermarking theory presented in the preceding sections with a well-known principle of speech perception, leading to a substantial improvement in theoretical capacity. From this we develop in Chapter 4 a practical watermarking scheme that is based on the above concept of replacing signal components (see Section 2.3) and using common speech coding and digital communications techniques.

The watermark power σ_W^2 in (2.1) can be interpreted as a mean squared error (MSE) host signal distortion constraint and essentially determines the watermark capacity. However, it has previously been shown that MSE is not a suitable distortion measure for

speech signals [41], as for example flipping the polarity of the signal results in a large MSE but in zero perceptual distortion. It is a long-known fact that, in particular for non-voiced speech (denoting all speech that is not voiced, comprising unvoiced speech and pauses) and blocks of short duration, the ear is insensitive to the signal's phase [42]. This effect is also exploited in audio coding for perceptual noise substitution [43]. Thus, instead of the MSE distortion (or watermark power σ_{W}^2) constraint, we propose to constrain the watermarked signal to have the same power spectral density (PSD) as the host signal (with power σ_{H}^2) but allow arbitrary modification (or replacement) of the phase. Compared to MSE this PSD constraint is far less restrictive and results in a watermark channel capacity

$$C_{\mathrm{W,\,Phase}} \approx \frac{1}{4}\log_2\left(\frac{\sigma_{\mathrm{H}}^2}{\sigma_{\mathrm{N}}^2}\right) + \frac{1}{4}\log_2\left(\frac{4\pi}{e}\right). \tag{2.2}$$

This high SNR approximation is derived in Section 3.4.2. Note that in contrast to (2.1) the capacity is no longer determined by the MSE distortion to channel noise ratio $\frac{\sigma_{\mathrm{W}}^2}{\sigma_{\mathrm{N}}^2}$, but by the host signal to channel noise ratio or SNR $\frac{\sigma_{\mathrm{H}}^2}{\sigma_{\mathrm{N}}^2}$, and the watermark signal has the same power as the host signal.

While watermarking in the phase domain is not a completely new idea, previously proposed methods either require the availability of the original host signal at the detector [44], are not suitable for narrowband speech [45, 46], or are restricted to relatively subtle phase modifications [14, 47, 48]. Also, the large theoretical watermark capacity given by (2.2) has not been recognized before.

To obtain a perceptually transparent and practical implementation of our phase embedding approach, we assume an autoregressive (AR) speech signal model and constrain the AR model spectrum and the temporal envelope to remain unchanged. There is a one-to-one mapping between the signal phase and the model's excitation, which allows us to modify the signal phase by modifying the excitation. Applying the concept of replacing certain signal components of Section 2.3, we exchange the model excitation by a watermark signal. We do so in non-voiced speech only, since in voiced speech certain phase spectrum changes are in fact audible [49, 50].

Chapter 3

Watermark Capacity in Speech

This chapter investigates the theoretical watermark capacity in speech given an additive white Gaussian noise transmission channel. Starting from the general capacity of the ideal Costa scheme, it shows that a large improvement in theoretical watermark capacity is possible if the application at hand allows watermarking in perceptually irrelevant signal components. Different ways to derive approximate and exact expressions for the watermark capacity when modulating the signal's DFT phase are presented and compared to an estimation of the capacity when modulating host signal frequency components that are perceptually masked. Parameters required for the experimental comparison of both methods are estimated from a database of ATC speech signals.

This chapter presents recent results that have not yet been published.

3.1. Introduction

In this chapter, we aim to estimate the watermark capacity in speech given an additive white Gaussian noise (AWGN) channel attack. Using well-established speech perception properties, we show that the watermark capacity in speech far exceeds the watermark capacity of conventional quantization-based watermarking. We achieve this improvement by proposing a clear distinction between watermarking in perceptually relevant signal components (or domains) and watermarking in perceptually irrelevant signal components.

Given that in the past most watermarking research was driven by the need of the multimedia content industry for copyright protection systems, many algorithms aim at a robustness against perceptual coding attacks. Since perceptual coding aims to remove perceptually irrelevant signal components, watermarking must occur in perceptually relevant components. However, there is a limit to how much perceptually relevant components can be altered without degrading perceptual quality to an unacceptable level. It is within this context where quantization-based watermarking and its capacity evolved, and its capacity is reviewed in Section 3.2.

In watermarking for analog legacy system enhancement, robustness against perceptual coding is not a concern. Besides watermarking in perceptually relevant components, watermarking in perceptually irrelevant signal components is possible, which opens a door to significant capacity improvements. A similar discussion can be found in [13], but otherwise the important distinction between watermarking in perceptually relevant or irrelevant components is seldom brought forward.

Given the aeronautical application presented in Chapter 6, we focus on watermarking in perceptually irrelevant components and derive the watermark capacities in speech on the basis of frequency masking (Section 3.3) and phase perception (Section 3.4). An experimental comparison of the different capacities is performed in Section 3.5.

In the remainder of this chapter, all capacities C are given in bits per sample (or bits per independent symbol, equivalently). Thus, the capacity specifies how many watermark bits can be embedded in an independent host signal sample or symbol. For a band-limited channel, the maximum number of independent symbols per second (the maximum symbol rate or Nyquist rate) is twice the channel bandwidth (e.g., [51]). Consequently, the maximum number of watermark bits transmitted per second is C times twice the channel bandwidth in Hz.

3.2. Watermarking Based on Ideal Costa Scheme

As a basis for comparison, we start with a brief review of the capacity of watermarking in perceptually relevant signal components and the Ideal Costa Scheme (ICS) [13]. The ICS, also termed 'dirty paper watermarking', 'watermarking with side information' or 'informed embedding' is the underlying principle for most quantization-based methods, and it is possible to derive an achievable upper bound for the watermark capacity.

In quantization-based watermarking the host signal is quantized using different vec-

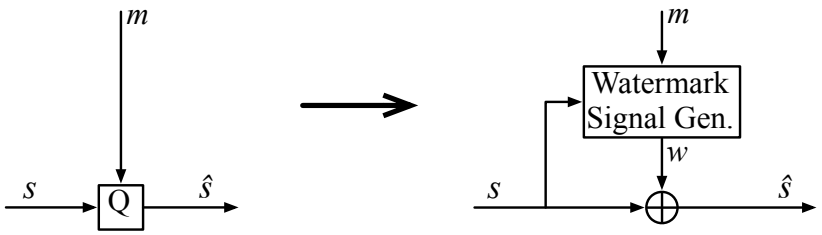


Figure 3.1.: Quantization-based watermarking as a host-signal-adaptive additive process.

tor quantizers that depend on the watermark data. Figure 3.1 depicts the quantization as an addition of a host signal dependent watermark, which has a power constraint σ_{W}^2 that corresponds to an MSE criterion for the distortion of the host signal. We review in the following the watermark capacity derivation given this MSE distortion criterion.

Assuming a white Gaussian host signal vector $\mathbf{s}$ of length L with variance σ_{H}^2, the watermarked signal $\hat{\mathbf{s}}$ must lie within an $(L-1)$-sphere W centered around $\mathbf{s}$ with volume

$$V_{\mathrm{W}} = \frac{\pi^{L/2} r_{\mathrm{W}}^L}{\Gamma(\frac{L}{2}+1)}$$

and radius $r_{\mathrm{W}} = \sqrt{L\sigma_{\mathrm{W}}^2}$. The Gamma function Γ is defined as

$$\Gamma(z+1) = z\Gamma(z) = z\int_0^\infty t^{z-1}e^{-t}\mathrm{d}t.$$

The transmitted signal is subject to AWGN with variance σ_{N}^2, and in high dimensions the received signal vector lies with high probability inside a hypersphere N with radius $r_{\mathrm{N}} = \sqrt{L\sigma_{\mathrm{N}}^2}$ centered at $\hat{\mathbf{s}}$ [52]. Consequently, all possible watermarked signals within W lie after transmission within a hypersphere Y with radius $r_{\mathrm{Y}} = \sqrt{L(\sigma_{\mathrm{W}}^2+\sigma_{\mathrm{N}}^2)}$ (see Figure 3.2). The number of distinguishable watermark messages given the channel noise is thus $V_{\mathrm{Y}}/V_{\mathrm{W}}$, which roughly means the number of 'noise spheres' N that fit into the 'allowed-distortion sphere' W. The watermark capacity evaluates to [51, 52, 13]

$$C_{\mathrm{ICS}} = \frac{1}{L}\log_2\left(\frac{V_{\mathrm{Y}}}{V_{\mathrm{W}}}\right) = \frac{1}{2}\log_2\left(1+\frac{\sigma_{\mathrm{W}}^2}{\sigma_{\mathrm{N}}^2}\right)\ \text{bit/sample} \tag{3.1}$$

and is inherently limited by the MSE distortion constraint.

3.3. Watermarking Based on Auditory Masking

Auditory masking describes the psychoacoustical principle that some sounds are not perceived in the temporal or spectral vicinity of other sounds [53]. In particular, frequency masking (or simultaneous masking) describes the effect that a signal is not

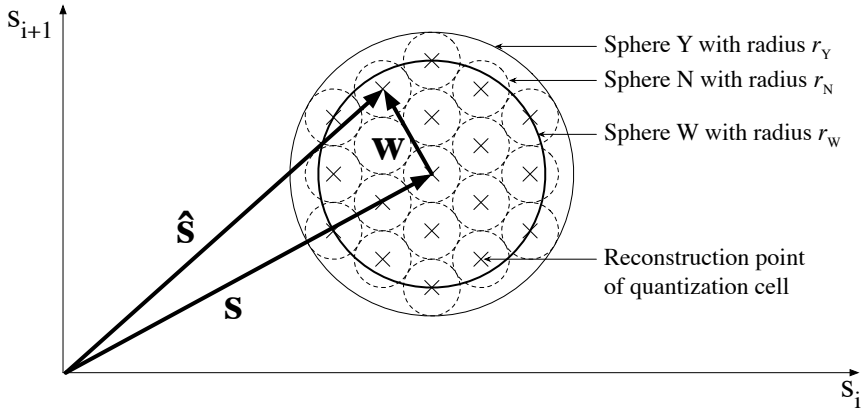


Figure 3.2.: Quantization-based watermarking and the Ideal Costa Scheme shown as sphere packing problem. Each reconstruction point (×) within the sphere W denotes a different watermark message.

audible in the presence of a simultaneous louder masker signal at nearby frequencies. Figure 3.3 shows the spectrum of a signal consisting of three sinusoids, and the masking threshold (or masking curve) derived using the van-de-Par masking model [54]. According to the model, sinusoidal components below the masking threshold are not audible because they are masked by the components above the masking threshold.

Perceptual models in general, and auditory masking in particular, are commonly used in practical audio watermarking schemes, and the principles are well understood [13, Ch. 8]. Besides more traditional approaches such as the spectral shaping of spread spectrum watermarks (e.g., [55, 28]), the principle of auditory masking is exploited either by varying the quantization step size or embedding strength in one way or the other (e.g., [56, 57]), or by removing masked components and inserting a watermark signal in replacement (e.g., [39]).

The remainder of this subsection derives an estimate of the theoretical watermark capacity that can be achieved based on frequency masking.

3.3.1. Theory

In quantization-based watermarking, the watermark signal corresponds to the quantization error induced by the watermarking process, and the watermark power σ_W^2 in (3.1) represents a MSE distortion criterion on the speech signal. If the watermark signal (or quantization error) is white, the watermark is perceived as background noise and the permissible level (the watermark power σ_W^2) is application-dependent. However, the watermark signal is inaudible if it is kept below the masking threshold. One can increase the watermark power to $\sigma_{W,mask}^2$, and thus the watermark capacity, by spectrally shaping the watermark signal according to the signal-dependent masking threshold. According to the masking model, the watermark signal is not audible as long as the spectral envelope of the watermark signal is at or below the masking threshold. The same approach is frequently applied in perceptual speech and audio coding: The

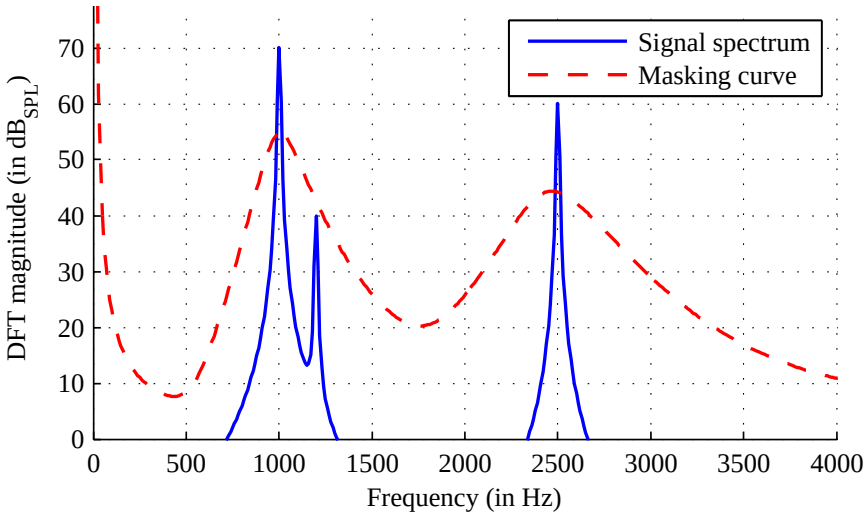


Figure 3.3.: Masking threshold for a signal consisting of three sinusoids, the second of them being masked and not audible.

quantization step-size is chosen adaptively such that the resulting quantization noise remains below the masking threshold [58, 59]. Using (3.1), the watermark capacity results in

$$C_{\text{Mask}} = \frac{1}{2}\log_2\left(1 + \frac{\sigma^2_{\text{W,mask}}}{\sigma^2_{\text{N}}}\right) \text{ bit/sample} \tag{3.2}$$

with typically $\sigma^2_{\text{W,mask}} \gg \sigma^2_{\text{W}}$.

3.3.2. Experimental Results

We estimate the permissible watermark power $\sigma^2_{\text{W,mask}}$ and, thus, the watermark capacity, with a small experiment. Using ten randomly selected utterances (one per speaker) of the ATCOSIM corpus (see Chapter 7), we calculate the masking thresholds in frames of 30 ms (overlapping by 50%). In a frequency band from 100 Hz to 8000 Hz we measure the average power σ^2_{H} of the speech signal as well as the permissible power $\sigma^2_{\text{W,mask}}$ of the watermark signal, with the spectral envelope of the watermark signal set to a factor κ below the signal-dependent masking threshold. The masking thresholds are calculated based on [54], using a listening-threshold-in-quiet as provided in [60], and a listening level of 82.5 dB$_{\text{SPL}}$, which corresponds to a comfortable speech signal level for normal listeners [61].

While the used masking model describes the masking of sinusoids by sinusoidal or white noise maskers, in this experiment the maskee is a wideband noise signal, which would not be entirely masked. To compensate for this shortcoming, the maskee signal is scaled to be a factor κ below the calculated masking threshold. The maximum value of κ at which the maskee signal is still inaudible was determined in an informal experiment. We generated a noise signal that is uncorrelated to the speech signal

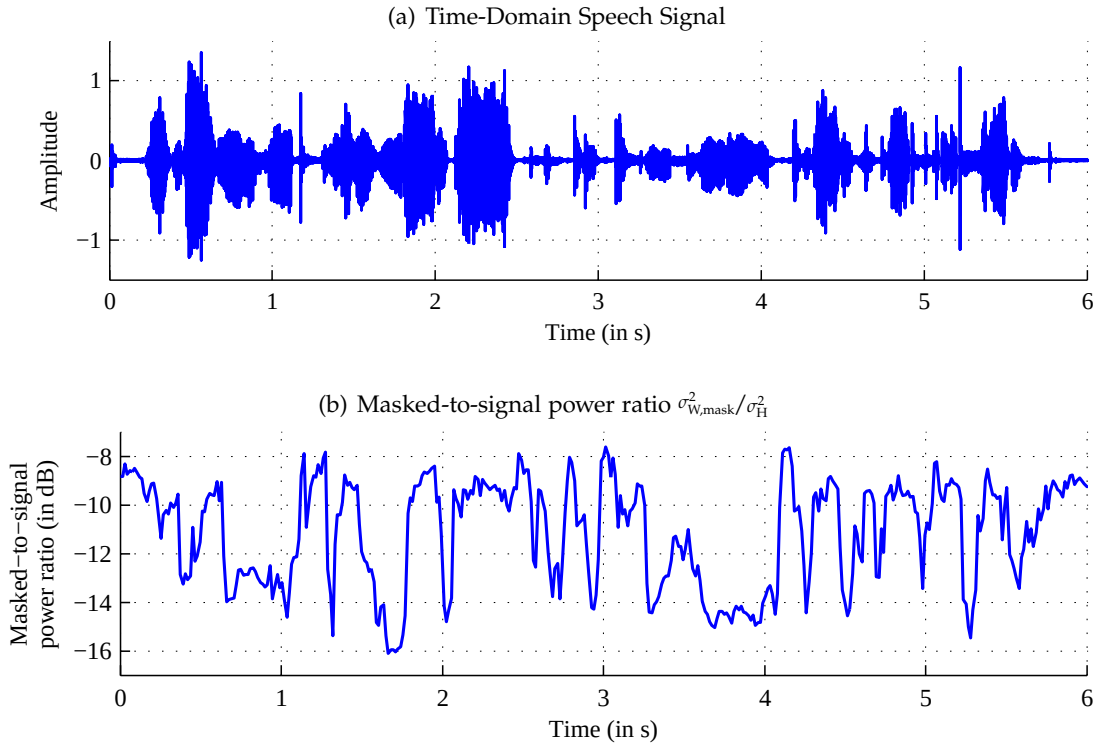


Figure 3.4.: Scaled masking threshold power to signal power ratio $\sigma^2_{W,mask}/\sigma^2_H$ using $\kappa = -12\,\text{dB}$ for the utterance "swissair eight zero six contact rhein radar one two seven decimal three seven".

and has a spectral envelope that corresponds to the masking threshold of the speech signal. This can be done by either setting the DFT magnitude of the watermark signal to the masking threshold and randomizing the DFT phase, or by generating in the DFT domain a complex white Gaussian noise with unit variance and multiplying this with the masking threshold. The resulting noise signal was added to the speech signal with varying gains κ, and the maximum value of κ at which the noise is still masked was found to be $\kappa \approx -12\,\text{dB}$, corresponding to one fourth in amplitude or one 16th in terms of power.

The masked-to-signal power ratio $\sigma^2_{W,mask}/\sigma^2_H$ for all frames of an utterance is shown in Figure 3.4. Averaged over all frames and utterances, the permissible watermark power is $\sigma^2_{W,mask} \approx \kappa\sigma^2_H$. The estimated average watermark capacity using (3.2) then results in

$$C_{\text{Mask}} \approx \frac{1}{2}\log_2\left(1+\kappa\frac{\sigma^2_H}{\sigma^2_N}\right) \approx \frac{1}{2}\log_2\left(1+\frac{1}{16}\frac{\sigma^2_H}{\sigma^2_N}\right). \tag{3.3}$$

Compared to (3.1), C_{Mask} does not depend on an MSE distortion criterion but is given by the transmission channel's signal-to-noise ratio (SNR) $\frac{\sigma^2_H}{\sigma^2_N}$.

3.4. Watermarking Based on Phase Modulation

We consider the case of watermarking by manipulating the phase spectrum of unvoiced speech, while keeping the power spectrum unchanged. This means that instead of an MSE distortion constraint we constrain $\hat{\mathbf{s}}$ to have the same power spectrum as $\mathbf{s}$, but allow arbitrary modification of the phase.

The following subsections present four different ways to derive the watermark capacity: First, two high SNR approximations are presented, one using the information theoretic definition of capacity, and one using sphere packing. Then, two exact solutions are derived, one based on the joint probability density function (PDF) between input and output phase angle, and one based on the conditional PDF of the complex output symbol given the complex input symbol.

3.4.1. Preliminaries

3.4.1.1. Definitions of Phase

The notion of a signal's phase is frequently used in literature. However, there is no standard definition, and the intended meaning of phase widely varies depending on the underlying concepts. In general, the phase of a signal is a quantity that is closely linked to a certain signal representation. Depending on the most suitable representation, we deal in the remainder of this thesis with three different 'phases', which are interrelated but nevertheless represent different quantities. The three phase definitions have in common that they represent a signal property that can be modified without modifying some sort of a spectral envelope of the signal.

AR Model Phase Primarily, we consider a particular realization of a white Gaussian excitation signal of an autoregressive (AR) speech signal model as defined in Section 4.1.1 as the phase of a signal. It was previously shown that for unvoiced speech the ear is not able to distinguish between different realizations of the Gaussian excitation process as long as the spectral and temporal envelope of the speech signal is maintained [42]. This notion of phase is used in our practical watermarking scheme of Chapter 4.

DFT Phase The DFT phase denotes the argument (or phase angle) of the complex coefficient of discrete Fourier transform (DFT) domain signal model as defined in Section 3.4.1.2. A mapping of an AR model phase to a DFT phase requires a short-term Fourier transform (STFT) of the signal using a well-defined transform size, window length and overlap. The following capacity derivations consider only a single DFT frame of length L of the speech signal.

MLT and ELT Phase In Section 3.4.8 modulated and extended lapped transform (MLT and ELT) signal representations are applied. In contrast to the DFT coefficients, the lapped transform coefficients are real and no direct notion of phase exists. Interpreting lapped transforms as filterbanks, we consider the variances of the subband

signals within a short window as the quantities that determine the spectral envelope, and consider the particular realizations of the subband signals as the phase of the signal. Thus, modifying the MLT or ELT phase of a signal means replacing the original subband signals by different subband signals with identical short-term power envelopes.

3.4.1.2. DFT Domain Signal Model

For all capacity derivations of Section 3.4 (except Section 3.4.8) we consider the discrete Fourier transform (DFT) of a single frame of length L of an unvoiced speech signal, and modify the DFT phase while keeping the DFT magnitude constant. We use a $\frac{1}{\sqrt{L}}$ weighting for the DFT and its inverse, which makes the DFT a unitary transform. Each frame is assumed to have L complex coefficients $\tilde{X}$, of which only $\frac{L}{2}$ are independent.[1] The tilde mark ($\tilde{}$) denotes complex variables, e.g., $\tilde{x} = re^{j\varphi} = a + jb$, whereas boldface letters denote real valued vectors, with $\mathbf{s}$, $\hat{\mathbf{s}}$ and $\mathbf{n}$ representing a single frame of the host signal, the watermarked signal and the channel noise, respectively. Considering the real and imaginary parts of $\tilde{X}$ as separate dimensions, the DFT representation has $2L$ real coefficients R, where the L odd-numbered dimensions represent the real part R_{rl} and and the L even-numbered dimensions represent the imaginary part R_{im} of the complex DFT coefficients $\tilde{X} = R_{\mathrm{rl}} + jR_{\mathrm{im}}$. Assuming a white host signal with variance σ_{H}^2, each complex DFT coefficient $\tilde{X}$ has variance σ_{H}^2, and each real coefficient R has variance $\frac{\sigma_{\mathrm{H}}^2}{2}$.

3.4.1.3. Watermarking Model

In order to facilitate theoretical watermark capacity derivations, we model the AR model phase replacement performed in [42] and Chapter 4 by replacing the phase of a DFT domain signal representation. An input phase angle Φ, which represents the watermark signal, is imposed onto a DFT coefficient of the host signal $\mathbf{s}$. While its original magnitude r_{H} is maintained, the phase angle is replaced by the watermark phase angle Φ, resulting in the watermarked DFT coefficient $\tilde{X} = r_{\mathrm{H}}e^{j\Phi}$. The watermarked signal is subject to AWGN $\mathbf{n}$ with DFT coefficient $\tilde{N}$, phase angle Ψ and variance σ_{N}^2, resulting in the noisy watermarked signal $\hat{\mathbf{s}} + \mathbf{n}$ with DFT coefficient $\tilde{Y} = \tilde{X} + \tilde{N}$ and output phase angle Θ.

3.4.1.4. PDFs of the Random Variables

In the following paragraphs, we derive the PDFs of the involved random variables (RV).

[1] For L even, the coefficients representing DC and Nyquist frequency are real-valued, and $\frac{L}{2} + 1$ coefficients are independent. For odd L, the DC coefficient is real, no Nyquist frequency coefficient exists, and $\frac{L+1}{2}$ coefficients are independent. As a simplification we assume in the following $\frac{L}{2}$ independent complex coefficients, which for large L leads to a negligible difference in the capacity estimates.

Input Phase Φ (scalar RV) Given the DFT domain signal model, the input phase angle is restricted to the interval $[0, 2\pi[$. To maximize its differential entropy h, which maximizes the watermark channel capacity C, Φ is assumed to be uniformly distributed with PDF

$$f_\Phi(\varphi) = \begin{cases} \frac{1}{2\pi} & 0 \leq \varphi < 2\pi \\ 0 & \text{else} \end{cases}$$

and differential entropy

$$h(\Phi) = \log_2(2\pi).$$

Noise Coefficient $\tilde{N}$ (complex RV) Each complex DFT coefficient $\tilde{N}_c$, the subscript c denoting Cartesian coordinates, of a complex zero-mean WGN signal with variance σ_N^2 is characterized by

$$\tilde{N} \sim \mathcal{CN}(0, \sigma_\mathrm{N}^2),$$

$$f_{\tilde{N}_c}(\tilde{x}) = \frac{1}{\pi\sigma_\mathrm{N}^2} \exp\left(-\frac{|\tilde{x}|^2}{\sigma_\mathrm{N}^2}\right) \quad \text{with} \quad |\tilde{x}|^2 = a^2 + b^2 = r^2, \qquad \text{and} \tag{3.4}$$

$$h(\tilde{N}) = \log_2(\pi e \sigma_\mathrm{N}^2).$$

But $f_{\tilde{N}}(\tilde{x})$ is really just a short-hand notation for the joint PDF $f_{N_a,N_b}(a,b)$ of the independent marginal densities

$$\begin{aligned} N_a &\sim \mathcal{N}(0, \tfrac{\sigma_\mathrm{N}^2}{2}) = \frac{1}{\sqrt{\pi\sigma_\mathrm{N}^2}} \exp\left(-\frac{a^2}{\sigma_\mathrm{N}^2}\right) \\ N_b &\sim \mathcal{N}(0, \tfrac{\sigma_\mathrm{N}^2}{2}) = \frac{1}{\sqrt{\pi\sigma_\mathrm{N}^2}} \exp\left(-\frac{b^2}{\sigma_\mathrm{N}^2}\right) \\ f_{\tilde{N}_c}(\tilde{x}) &= f_{N_a,N_b}(a,b) = f_{N_a}(a) f_{N_b}(b) \end{aligned}$$

and $h(N_a) = h(N_b) = \frac{1}{2}\log_2(\pi e \sigma_\mathrm{N}^2)$. Note that $h(\tilde{N}) = h(N_a) + h(N_b)$, since the joint differential entropy is the sum of the differential entropies if the variables are independent.

The density $f_{\tilde{N}_c}(\tilde{x})$ is not the density in polar coordinates, i.e. $f_{\tilde{N}_p}(r,\varphi) \neq f_{\tilde{N}_c}(\tilde{x})$, because a transformation of variables is required. This results in [62, p. 201 ff.][63]

$$f_{N_\varphi}(\varphi) = \begin{cases} \frac{1}{2\pi} & 0 \leq \varphi < 2\pi \\ 0 & \text{else} \end{cases}$$

$$f_{N_r}(r) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) = \frac{2r}{\sigma_\mathrm{N}^2} \exp\left(-\frac{r^2}{\sigma_\mathrm{N}^2}\right) \tag{3.5}$$

where $f_{N_r}(r)$ is the Rayleigh distribution with parameter σ (in our case with $\sigma^2 = \frac{\sigma_{\tilde{N}}^2}{2}$) and mean and variance

$$\begin{aligned} \mu_{N_r} &= \sigma\sqrt{\frac{\pi}{2}} = \frac{\sigma_{\mathrm{N}}}{2}\sqrt{\pi} \\ \sigma_{N_r}^2 &= (2 - \frac{\pi}{2})\sigma^2 = (1 - \frac{\pi}{4})\sigma_{\mathrm{N}}^2. \end{aligned}$$

In this special case, comparing (3.4) and (3.5) results in $f_{N_r}(r) = 2\pi r \cdot f_{\tilde{N}_c}(\tilde{x})$. Also, N_r and N_φ are independent, and their joint distribution is [62, p. 258]

$$f_{\tilde{N}_p}(\tilde{x}) = f_{N_r,N_\varphi}(r,\varphi) = f_{N_r}(r) f_{N_\varphi}(\varphi) = \frac{r}{\pi\sigma_{\mathrm{N}}^2} \exp\left(-\frac{r^2}{\sigma_{\mathrm{N}}^2}\right) \tag{3.6}$$

and $f_{\tilde{N}_p}(\tilde{x}) = r \cdot f_{\tilde{N}_c}(r)$. This polar representation can be derived from (3.4) by considering the polar coordinates (r, φ) as a vector function (or transform) of the Cartesian coordinates (a, b), with

$$\begin{aligned} r &= \sqrt{a^2 + b^2} \\ \varphi &= \arctan\tfrac{b}{a} \end{aligned}$$

and the inverse transformation

$$\begin{aligned} a &= r\cos\varphi \\ b &= r\sin\varphi. \end{aligned}$$

The Jacobian determinant J of the transform is

$$J_N = \frac{\partial(a,b)}{\partial(r,\varphi)} = \begin{vmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{vmatrix} = r$$

and the joint density is

$$f_{N_r,N_\varphi}(r,\varphi) = J_N \cdot f_{N_a,N_b}(r\cos\varphi, r\sin\varphi) = \frac{r}{\pi\sigma_{\mathrm{N}}^2} \exp\left(-\frac{r^2}{\sigma_{\mathrm{N}}^2}\right).$$

Signal Coefficient $\tilde{X}$ (complex RV) We consider a single complex DFT coefficient

$$\tilde{X} = r_{\mathrm{H}} e^{j\Phi}$$

with r_{H} being a deterministic constant. In polar coordinates, the PDFs of the independent RVs X_φ and X_r are

$$\begin{aligned} f_{X_\varphi}(\varphi) &= \begin{cases} \frac{1}{2\pi} & 0 \le \varphi < 2\pi \\ 0 & \text{else} \end{cases} \\ f_{X_r}(r) &= \delta(r - r_{\mathrm{H}}) \\ f_{X_r,X_\varphi}(r,\varphi) &= \begin{cases} \frac{1}{2\pi}\delta(r - r_{\mathrm{H}}) & 0 \le \varphi < 2\pi \\ 0 & \text{else} \end{cases} \end{aligned}$$

and describe a two-dimensional circularly symmetric distribution [63].

With the transform

$$\begin{aligned} a &= r\cos\varphi \\ b &= r\sin\varphi \end{aligned}$$

and the inverse transform

$$\begin{aligned} r &= \sqrt{a^2+b^2} \\ \varphi &= \arctan\tfrac{b}{a} \end{aligned}$$

and its Jacobian

$$J_X = \frac{\partial(r,\varphi)}{\partial(a,b)} = \begin{vmatrix} \frac{\partial r}{\partial a} & \frac{\partial r}{\partial b} \\ \frac{\partial \varphi}{\partial a} & \frac{\partial \varphi}{\partial b} \end{vmatrix} = \frac{1}{\sqrt{a^2+b^2}}$$

the joint density in Cartesian coordinates is

$$\begin{aligned} f_{X_a,X_b}(a,b) = J_X \cdot f_{X_r,X_\varphi}\left(\sqrt{a^2+b^2}, \tan\tfrac{b}{a}\right) = \frac{1}{2\pi\sqrt{a^2+b^2}}\delta\left(\sqrt{a^2+b^2}-r_{\mathrm{H}}\right) = \\ = \frac{1}{2\pi r_{\mathrm{H}}}\delta\left(\sqrt{a^2+b^2}-r_{\mathrm{H}}\right) \end{aligned} \tag{3.7}$$

and the dependent marginal densities in Cartesian coordinates are

$$f_{X_a}(a) = \begin{cases} \frac{1}{\pi r_{\mathrm{H}}}\left(1-(\frac{a}{r_{\mathrm{H}}})^2\right)^{-\frac{1}{2}} & \text{for } |a| \le r_{\mathrm{H}} \\ 0 & \text{else} \end{cases} \tag{3.8}$$

$$f_{X_b}(b) = \begin{cases} \frac{1}{\pi r_{\mathrm{H}}}\left(1-(\frac{b}{r_{\mathrm{H}}})^2\right)^{-\frac{1}{2}} & \text{for } |b| \le r_{\mathrm{H}} \\ 0 & \text{else.} \end{cases}$$

The marginal densities are derived as follows. The arc length Δc within an infinitesimally small strip of width Δa as shown in Figure 3.5 can be approximated by the length of the tangent within the strip. Using

$$\sin\beta = \frac{A}{r_{\mathrm{H}}} = \frac{f}{\Delta c} = \frac{\sqrt{c^2-(\Delta a)^2}}{\Delta c}$$

the arc length results in

$$\Delta c = \left(1-\left(\frac{A}{r_{\mathrm{H}}}\right)^2\right)^{-\frac{1}{2}}\Delta a.$$

The probability $\mathcal{P}$ that a point lies within the strip Δa is the arc length Δc times the density $\frac{1}{2\pi r_{\mathrm{H}}}$ of the arc, times a factor of 2 to also account for the lower semicircle,

$$\mathcal{P} = \frac{1}{\pi r_{\mathrm{H}}}\Delta c,$$

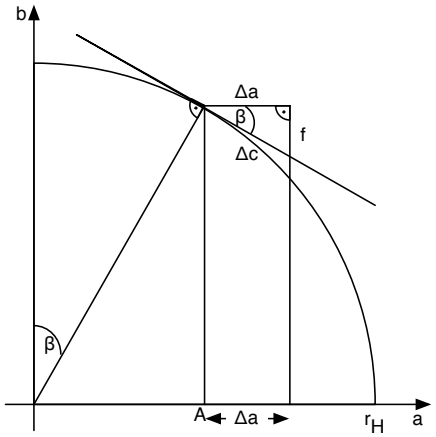


Figure 3.5.: The joint density $f_{X_a,X_b}(a,b)$ integrated over an interval Δa.

and the probability density $\frac{\mathcal{P}}{\Delta a}$ gives in the limit $\Delta a \to 0$ the marginal PDF

$$f_{X_a}(a) = \int_{b=-\infty}^{\infty} f_{X_a,X_b}(a,b)\mathrm{d}b = \begin{cases} \frac{1}{\pi r_\mathrm{H}} \left(1 - \left(\frac{a}{r_\mathrm{H}}\right)^2\right)^{-\frac{1}{2}} & \text{for } |a| \leq r_\mathrm{H} \\ 0 & \text{else.} \end{cases}$$

The marginal density integrates to 1 since

$$\int_{a=-\infty}^{\infty} f_{X_a}(a)\mathrm{d}a = \frac{1}{\pi r_\mathrm{H}} \int_{a=-r_\mathrm{H}}^{r_\mathrm{H}} \left(1 - \left(\frac{a}{r_\mathrm{H}}\right)^2\right)^{-\frac{1}{2}} \mathrm{d}a = \\ = \frac{1}{\pi r_\mathrm{H}} \left[r_\mathrm{H} \arcsin\left(\frac{a}{r_\mathrm{H}}\right)\right]_{-r_\mathrm{H}}^{r_\mathrm{H}} = 1.$$

Noisy Signal Coefficient $\tilde{Y}$ (complex RV) The noise-corrupted watermarked coefficient $\tilde{Y}$ is a sum of two independent complex RV, with

$$\tilde{Y} = \tilde{X} + \tilde{N} = r_Y e^{j\Theta}, \qquad \text{and}$$
$$f_{\tilde{Y}}(\tilde{x}) = f_{\tilde{Y}}(r) = f_{\tilde{X}}(\tilde{x}) \ast\ast f_{\tilde{N}}(\tilde{x})$$

where $\ast\ast$ denotes 2D-convolution. In Cartesian coordinates the summation of the two complex RVs can be carried out on a by component basis, with

$$\begin{aligned} Y_a &= X_a + N_a \\ Y_b &= X_b + N_b. \end{aligned}$$

Using (3.8), the density of the first sum is

$$f_{Y_a}(a) = f_{X_a}(a) * f_{N_a}(a) = \int_{m=-\infty}^{\infty} f_{X_a}(m) f_{N_a}(a-m)\mathrm{d}m = \\ = \int_{m=-\infty}^{\infty} f_{X_a}(a-m) f_{N_a}(m)\mathrm{d}m,$$

which evaluates to

$$f_{Y_a}(a) = \frac{1}{\pi r_\mathrm{H}} \frac{1}{\sqrt{\pi\sigma_\mathrm{N}^2}} \int_m \left(1 - \left(\frac{m}{r_\mathrm{H}}\right)^2\right)^{-\frac{1}{2}} \exp\left(-\frac{(m-a)^2}{\sigma_\mathrm{N}^2}\right) \mathrm{d}m.$$

No analytic solution to this integral is known. The expression for $f_{Y_b}(b)$ is equivalent.

Since X_a and X_b are dependent, Y_a and Y_b are also dependent, and the joint density is not the product of the marginal densities. It appears difficult to obtain closed-form expressions for the joint and conditional PDFs of $\tilde{X}$ and $\tilde{Y}$. However, as shown in Section 3.4.5 and Section 3.4.6, the conditional PDF $f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x})$ can be easily stated given the circular arrangement of $\tilde{X}$ and $\tilde{Y}$.

Output Phase Θ (scalar RV) The output phase Θ of the observed watermarked signal $\tilde{Y}$ is

$$\Theta = \arg(\tilde{Y}) = \arctan\frac{Y_b}{Y_a} = Y_\varphi,$$

and is again uniformly distributed, because the input phase is uniformly distributed and the channel noise is Gaussian. Thus,

$$f_\Theta(\varphi) = \begin{cases} \frac{1}{2\pi} & 0 \le \varphi < 2\pi \\ 0 & \text{else} \end{cases} \qquad \text{and}$$

$$h(\Theta) = \log_2(2\pi).$$

3.4.1.5. Approaches to Capacity Calculation

Given the DFT-based signal and watermarking model, two watermark capacities can be calculated, one based on the channel input/output phase angles Φ and Θ, and one based on the channel input/output complex symbols $\tilde{X}$ and $\tilde{Y}$. Even though information is only embedded in the phase angle Φ of $\tilde{X}$, the two capacities are not identical, since in the later case the watermark detector can also respond to the amplitude of the received signal.

The capacity is defined as the mutual information between the channel input and the channel output, maximized over all possible input distributions, and

$$C = \frac{1}{2} \sup_{f_{\tilde{X}}(\tilde{x})} I(\tilde{X}, \tilde{Y}) \quad \text{or} \quad C = \frac{1}{2} \sup_{f_\Phi(\varphi)} I(\Phi, \Theta). \tag{3.9}$$

The additional factor $\frac{1}{2}$ compared to the usual definition results from only half of the complex DFT coefficients $\tilde{X}$ being independent. We calculate the capacity based on:

1. the sphere packing analogy of Section 3.2 (used in Section 3.4.3);
2. the difference in differential entropy between $\tilde{Y}$ and $\tilde{N}$ (used in Section 3.4.2);
3. the formal definition of I based on the input's PDF and the conditional PDF of the output given the input (used in Section 3.4.4, 3.4.5 and 3.4.6).

3.4.1.6. Colored Host Signal Spectrum

While most of the following derivations assume a white host signal spectrum, the extension to a colored host signal is straightforward and shown for the two high SNR approximations of Section 3.4.2 and 3.4.3. The capacity can be calculated individually for each DFT channel, and the overall capacity is the sum of the subchannel capacities. We introduce band power weights p_i, with

$$\sum_{i=1}^{L/2} p_i = \frac{L}{2} \tag{3.10}$$

, which effectively scale the SNR of the i'th subchannel from $\frac{\sigma_H^2}{\sigma_N^2}$ to $p_i \frac{\sigma_H^2}{\sigma_N^2}$. Using a white host signal, $\forall i : p_i = 1$. The discrete band power weights p_i could also be considered as the samples of a scaled power spectral density $p(f)$ of the host signal. Substituting L by f_s and considering (3.10) as Riemann sum results equivalently in

$$\int_{f=0}^{f_s/2} p(f) \mathrm{d}f = \frac{f_s}{2}. \tag{3.11}$$

For a white host signal, $p(f) \equiv 1$ as in the discrete case.

3.4.2. High SNR Capacity Approximation Using Mutual Information

We first consider the complex DFT coefficient $\tilde{X}$ of a white host signal with power σ_H^2. The Gaussian channel noise $\tilde{N}$ with power σ_N^2 changes both the amplitude and the phase of the received signal. The 'displacement' N of $\tilde{X}$ in angular direction caused by $\tilde{N}$ corresponds to a phase change $\tan(\varphi) = \frac{N}{\sigma_H}$. In the case of high SNR, the noise component of $\tilde{N}$ in angular direction is a scalar Gaussian random variable N with power $\frac{\sigma_N^2}{2}$, and we can also assume $\varphi \approx \tan(\varphi)$. The phase noise Ψ induced by the channel is then also Gaussian, with

$$\Psi \sim \mathcal{N}\left(0, \frac{\sigma_N^2/2}{\sigma_H^2}\right).$$

Given a transmitted phase Φ, the observed phase Θ is then $\Theta = \Phi + \Psi$, wrapped into $0 \ldots 2\pi$.

The channel capacity $C_{\text{Phase, MI-angle}}$ is the maximum mutual information $I(\Phi;\Theta)$ over all distributions of Φ. The mutual information is maximized when Φ (and consequently

also Θ since Ψ is Gaussian) is uniformly distributed between 0 and 2π as defined above. The maximum mutual information is then given by

$$\begin{aligned} I(\Phi;\Theta) &= h(\Theta) - h(\Theta|\Phi) \\ &= h(\Theta) - h(\Theta - \Phi|\Phi) = h(\Theta) - h(\Psi|\Phi) = h(\Theta) - h(\Psi) \end{aligned}$$

since Φ and Ψ are independent. The mutual information $I(\Phi,\Theta)$ between the received phase Θ and the transmitted phase Φ then evaluates to

$$\begin{aligned} I(\Phi,\Theta) &= h(\Theta) - h(\Psi) = \log_2(2\pi) - \frac{1}{2}\log_2(2\pi e \frac{\sigma_\mathrm{N}^2}{2\sigma_\mathrm{H}^2}) \\ &= \frac{1}{2}\log_2\left(\frac{\sigma_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\right) + \frac{1}{2}\log_2\left(\frac{4\pi}{e}\right), \end{aligned}$$

and with only half of the complex coefficients $\tilde{X}$ being independent we obtain

$$C_{\text{Phase, MI-angle, high-SNR}} = \frac{1}{2} I(\Phi,\Theta) = \frac{1}{4}\log_2\left(\frac{\sigma_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\right) + \frac{1}{4}\log_2\left(\frac{4\pi}{e}\right), \tag{3.12}$$

with $\frac{1}{4}\log_2\left(\frac{4\pi}{e}\right) \approx 0.55\,\text{bit}$. The subscript $_{\text{Phase, MI-angle, high-SNR}}$ denotes a high-SNR approximation of the phase modulation watermark capacity derived using the mutual information (MI) between the input and output phase angles. The same result is obtained elsewhere for phase shift keying (PSK) in AWGN, but requiring a longer proof [64].

For a non-white host signal with band power weights p_i, the capacity results in

$$\begin{aligned} C_{\text{Phase, MI-angle, high-SNR, colored}} &= \frac{1}{2L}\log_2\left(\prod_{i=1}^{L/2} p_i \frac{\sigma_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\right) + \frac{1}{4}\log_2\left(\frac{4\pi}{e}\right) = \\ &= \frac{1}{4}\log_2\left(\frac{\sigma_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\right) + \frac{1}{4}\log_2\left(\frac{4\pi}{e}\right) + \frac{1}{2L}\sum_{i=1}^{L/2}\log_2\left(p_i\right). \end{aligned}$$

The factor $\frac{1}{2L}$ results from the factor $\frac{1}{4}$ of (3.12) and a factor $\frac{1}{L/2}$ given by the number of product or summation terms. Using the scaled host signal PSD $p(f)$ of Section 3.4.1.6 instead of the discrete band power weights p_i results in

$$\begin{aligned} &C_{\text{Phase, MI-angle, high-SNR, colored}} = \\ &\qquad = \frac{1}{4}\log_2\left(\frac{\sigma_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\right) + \frac{1}{4}\log_2\left(\frac{4\pi}{e}\right) + \frac{1}{2f_s}\int_{f=0}^{f_s/2}\log_2\left(p(f)\right)\mathrm{d}f. \end{aligned}$$

3.4.3. High SNR Capacity Approximation Using Sphere Packing

An estimate of the watermark capacity can also be obtained using the same line of arguments as in Section 3.2. Instead of the watermarked signal $\hat{\mathbf{s}}$ being constrained to a hypersphere volume around $\mathbf{s}$, it is now constrained to lie on an $\frac{L}{2}$-dimensional

manifold which contains all points in the L-dimensional space that have the same power spectrum as $\mathbf{s}$.

In the DFT domain, the power spectrum constraint corresponds to a fixed amplitude of the complex DFT coefficients $\tilde{X}$. In terms of the real and imaginary coefficients R_{rl} and R_{im} this means that for every 2-dimensional subspace of corresponding $(R_{\mathrm{rl}}; R_{\mathrm{im}})$ the watermarked signal must lie on a circle with radius $r_{\mathrm{H}} = \sqrt{p_i \sigma_{\mathrm{H}}^2}$, with p_i as defined above (for a white signal $p_i \equiv 1$). The watermarked signal $\hat{\mathbf{s}}$ must thus lie on the $\frac{L}{2}$-dimensional manifold that is defined by these circles. The surface of this manifold is

$$A_{\mathrm{H}} = \prod_{i=1}^{L/2} 2\pi r_{\mathrm{H}}. \tag{3.13}$$

The transmitted signal is subject to AWGN with variance σ_{N}^2. We need to determine the constellation of the observed signal and the number of distinguishable watermark messages that fit onto the manifold.[2] Averaged over all dimensions and given the circular symmetry of the distributions, the noise can be split up into two components of equal power $\frac{\sigma_{\mathrm{N}}^2}{2}$, one being orthogonal to the above manifold, and one being parallel to the manifold. Separating the noise into a parallel and an orthogonal component requires that the manifold is locally flat with respect to the size of the noise sphere, which is equivalent to a high SNR assumption. The orthogonal component effectively increases the radii of the circles from r_{H} to $r_{\mathrm{Y}} = \sqrt{p_i \sigma_{\mathrm{H}}^2 + \frac{\sigma_{\mathrm{N}}^2}{2}}$.[3] The noise component parallel to the manifold (with power $\frac{\sigma_{\mathrm{N}}^2}{2}$) determines the number of distinguishable watermark messages. The 'size' of a watermark cell on the $\frac{L}{2}$ dimensional manifold is the volume V_{N} of the $\frac{L}{2}$-dimensional hypersphere with radius $r_{\mathrm{N}} = \sqrt{\frac{L}{2}\frac{\sigma_{\mathrm{N}}^2}{2}}$, and

$$V_{\mathrm{N}} = \frac{(\pi \frac{L}{2} \frac{\sigma_{\mathrm{N}}^2}{2})^{L/4}}{\Gamma(\frac{L}{4}+1)}. \tag{3.14}$$

The capacity is

$$C_{\text{Phase, SP, high-SNR}} = \frac{1}{L} \log_2 \left(\frac{A_{\mathrm{Y}}}{V_{\mathrm{N}}} \right) \tag{3.15}$$

which evaluates for a white host signal ($p_i \equiv 1$) to

$$C_{\text{Phase, SP, high-SNR}} = \frac{1}{4} \log_2 \left(\frac{1}{2} + \frac{\sigma_{\mathrm{H}}^2}{\sigma_{\mathrm{N}}^2} \right) + \frac{1}{4} \log_2 (16\pi) + \frac{1}{L} \log_2 \left(\frac{\Gamma(\frac{L}{4}+1)}{L^{L/4}} \right). \tag{3.16}$$

The subscript $_{\mathrm{SP}}$ denotes a watermark capacity derived using the sphere packing approach. For large L and using Sterling's approximation, $\Gamma(\frac{L}{4}+1)$ can be expressed

[2] The projection of the L-dimensional noise (with spherical normal PDF) onto the $\frac{L}{2}$-dimensional subspace would result in the volume of the $\frac{L}{2}$-dimensional hypersphere with radius $r_{\mathrm{N}} = \sqrt{\frac{L}{2}\sigma_{\mathrm{N}}^2}$. A projection of the noise onto a $\frac{L}{2}$ dimensional subspace (line, plane, hyperplane, ...) is something else than a projection onto a curved manifold.

[3] This also reflects the fact that the channel noise changes the power spectrum of the received signal.

as

$$\Gamma(\tfrac{L}{4}+1) \approx \sqrt{\frac{\pi L}{2}}\left(\frac{L}{4e}\right)^{\frac{L}{4}}$$

and (3.16) evaluates to

$$C_{\text{Phase, SP, high-SNR}} = \frac{1}{4}\log_2\left(\frac{1}{2}+\frac{\sigma_{\text{H}}^2}{\sigma_{\text{N}}^2}\right)+\frac{1}{4}\log_2\left(\frac{4\pi}{e}\right)+\frac{1}{2L}\left[\log_2(\pi L)+1\right]. \quad (3.17)$$

Sterling's approximation is asymptotically accurate, and numerical simulations show that for $L > 70$ the error in C induced by the Sterling approximation is below 10^{-4}.

Again for large L, the last term on the right-hand side in (3.17) converges to zero and (3.16) simplifies to

$$C_{\text{Phase, SP, high-SNR}} = \frac{1}{4}\log_2\left(\frac{1}{2}+\frac{\sigma_{\text{H}}^2}{\sigma_{\text{N}}^2}\right)+\frac{1}{4}\log_2\left(\frac{4\pi}{e}\right). \quad (3.18)$$

The difference between (3.17) and (3.18) is smaller than 10^{-2} bit for $L > 480$. Given $\frac{\sigma_{\text{H}}^2}{\sigma_{\text{N}}^2} \gg 1$, the augend $\frac{1}{2}$ in (3.18) is negligible and the result is identical to (3.12).

For a non-white host signal, that means $\exists i : p_i \neq 1$, substituting (3.13) and (3.14) into (3.15) results in

$$C_{\text{Phase, SP, high-SNR, colored}} = \frac{1}{L}\log_2\left[\frac{\prod_{i=1}^{L/2} 2\pi\sqrt{p_i\sigma_{\text{H}}^2+\frac{\sigma_{\text{N}}^2}{2}}}{\left(\pi\frac{L}{2}\frac{\sigma_{\text{N}}^2}{2}\right)^{\frac{L}{4}}}\cdot\Gamma\left(\tfrac{L}{4}+1\right)\right] =$$

$$= \frac{1}{2L}\log_2\left(\prod_{i=1}^{L/2}\frac{1}{2}+p_i\frac{\sigma_{\text{H}}^2}{\sigma_{\text{N}}^2}\right)+\frac{1}{4}\log_2(16\pi)+\frac{1}{L}\log_2\left(\frac{\Gamma(\frac{L}{4}+1)}{L^{L/4}}\right) \quad (3.19)$$

which for large L again using Sterling's approximation simplifies to

$$C_{\text{Phase, SP, high-SNR, colored}} = \frac{1}{2L}\sum_{i=1}^{L/2}\log_2\left(\frac{1}{2}+p_i\frac{\sigma_{\text{H}}^2}{\sigma_{\text{N}}^2}\right)+\frac{1}{4}\log_2\left(\frac{4\pi}{e}\right). \quad (3.20)$$

Using the scaled host signal PSD $p(f)$ of Section 3.4.1.6 results in

$$C_{\text{Phase, SP, high-SNR, colored}} = \frac{1}{2f_s}\int_{f=0}^{f_s/2}\log_2\left(\frac{1}{2}+p(f)\frac{\sigma_{\text{H}}^2}{\sigma_{\text{N}}^2}\right)\mathrm{d}f+\frac{1}{4}\log_2\left(\frac{4\pi}{e}\right).$$

3.4.4. Exact Capacity Using Input/Output Phase Angles

Assuming a *deterministic* input signal $\tilde{X} = r_{\text{H}}e^{j0}$, the joint PDF of $\tilde{Y}$ is [65, p. 413]

$$f_{\tilde{Y}_p}(\tilde{x})|_{\tilde{X}=r_{\text{H}}e^{j0}} = f_{Y_r,Y_\varphi}(r,\varphi)|_{\tilde{X}=r_{\text{H}}e^{j0}} = \frac{r}{\pi\sigma_{\text{N}}^2}\exp\left(-\frac{r^2+r_{\text{H}}^2-2rr_{\text{H}}\cos(\varphi)}{\sigma_{\text{N}}^2}\right)$$

and the marginal PDF $f_{Y_\varphi}(\varphi)|_{\tilde{X}=r_\mathrm{H}e^{j0}}$ is [66]

$$f_{Y_\varphi}(\varphi)|_{\tilde{X}=r_\mathrm{H}e^{j0}} = \int_{r=0}^{\infty} f_{Y_r,Y_\varphi}(r,\varphi)\mathrm{d}r = \\ = \frac{1}{2\pi}\exp\left(-\frac{r_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\right) + \frac{r_\mathrm{H}\cos(\varphi)}{\sqrt{4\pi\sigma_\mathrm{N}^2}}\exp\left(-\frac{r_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\sin^2(\varphi)\right)\operatorname{erfc}\left(-\frac{r_\mathrm{H}}{\sigma_\mathrm{N}}\cos(\varphi)\right).$$

Provided the circular symmetry of the arrangement, changing the assumption of a deterministic $\tilde{X} = r_\mathrm{H}e^{j0}$ to a deterministic $\tilde{X} = r_\mathrm{H}e^{j\varphi'}$ corresponds to replacing (φ) by $(\varphi - \varphi')$ in the above two equations and represents a simple rotation. We also note that f_{Y_φ} (which is the same as f_Θ) given a deterministic input φ' is the conditional PDF $f_{\Theta|\Phi}(\varphi|\varphi')$ and write

$$f_{\Theta|\Phi}(\varphi|\varphi') = \frac{1}{2\pi}\exp\left(-\frac{r_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\right) + \\ + \frac{r_\mathrm{H}\cos(\varphi-\varphi')}{\sqrt{4\pi\sigma_\mathrm{N}^2}}\exp\left(-\frac{r_\mathrm{H}^2}{\sigma_\mathrm{N}^2}\sin^2(\varphi-\varphi')\right)\operatorname{erfc}\left(-\frac{r_\mathrm{H}}{\sigma_\mathrm{N}}\cos(\varphi-\varphi')\right).$$

With $f_{\Theta|\Phi}(\varphi|\varphi')\cdot f_\Phi(\varphi') = f_{\Phi,\Theta}(\varphi,\varphi')$ we also obtain the joint PDF

$$f_{\Phi,\Theta}(\varphi,\varphi') = \begin{cases} \frac{1}{2\pi}f_{\Theta|\Phi}(\varphi|\varphi') & \text{if } 0 \le \varphi,\varphi' < 2\pi \\ 0 & \text{else.} \end{cases}$$

The mutual information $I(\Phi,\Theta)$ can then be calculated using [64, p. 244]

$$I(\Phi,\Theta) = \int_{\varphi=0}^{2\pi}\int_{\varphi'=0}^{2\pi} f_{\Phi,\Theta}(\varphi',\varphi)\log_2\frac{f_{\Phi,\Theta}(\varphi',\varphi)}{f_\Phi(\varphi')f_\Theta(\varphi)}\mathrm{d}\varphi'\mathrm{d}\varphi.$$

Using (3.9) with only half of the DFT coefficients being independent, the capacity evaluates to

$$C_{\text{Phase, MI-angle}} = \frac{1}{2}I(\Phi,\Theta). \tag{3.21}$$

The required integrations can be carried out numerically. A similar derivation leading to the same result can be found in [66].

3.4.5. Exact Capacity Using Discrete Input/Output Symbols

Assume the input phase Φ consists of M discrete and uniformly spaced values. This corresponds to M-ary phase shift keying (PSK) using a constellation with M symbols, and the capacity is derived in [67, 68]. We briefly restate this derivation since it is the basis for the capacity derivation in Section 3.4.6.

We assume M discrete complex channel inputs $\tilde{x}_m$. The complex output sample $\tilde{y}$ has the conditional PDF

$$f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_m) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{|\tilde{y}-\tilde{x}_m|^2}{2\sigma^2}\right)$$

with $\sigma^2 = \frac{\sigma_{\mathrm{N}}^2}{2}$ and $\tilde{x}_m = r_{\mathrm{H}} \exp\left(j\frac{2\pi m}{M}\right)$. The capacity of the discrete-input continuous-output memoryless channel is

$$C_{\text{Phase, MI-c.sym., discrete}} = \frac{1}{2} \sup_{f_{\tilde{X}}(\tilde{x})} \sum_{m=0}^{M-1} f_{\tilde{X}}(\tilde{x}_m) \iint_{\tilde{y}=-\infty}^{\infty} f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_m) \cdot \log_2 \left\{ \frac{f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_m)}{\sum_{n=0}^{M-1} f_{\tilde{X}}(\tilde{x}_n) f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_n)} \right\} \mathrm{d}\tilde{y},$$

again with only half of the DFT coefficients $\tilde{X}$ being independent and where $\iint_{\tilde{y}=-\infty}^{\infty} \ldots \mathrm{d}\tilde{y}$ denotes an improper double integral over the two dimensions of the two-dimensional variable $\tilde{y}$, i.e., $\iint_{\tilde{y}=-\infty}^{\infty} \ldots \mathrm{d}\tilde{y} = \int_{y_a=-\infty}^{\infty} \int_{y_b=-\infty}^{\infty} \ldots \mathrm{d}y_b \mathrm{d}y_a$. The subscript $_{\text{Phase, MI-c.sym., discrete}}$ denotes a phase modulation watermark capacity derived using the mutual information (MI) between a set of discrete complex input symbols and the complex output symbols. Given the desired signal constellation, the capacity is maximum if the channel inputs $\tilde{x}_m$ are equiprobable with probability $f_{\tilde{X}}(\tilde{x}_m) = \frac{1}{M}$ and the supremum in the above term can be omitted. We further substitute $\tilde{Y} := \tilde{W} + \tilde{X}_m$ with $\tilde{W} \sim \mathcal{CN}(0, \sigma_{\mathrm{N}}^2)$, resulting in

$$f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_m) = f_{\tilde{W}}(\tilde{w}) = \frac{1}{\pi\sigma_{\mathrm{N}}^2} \exp\left(-\frac{|\tilde{w}|^2}{\sigma_{\mathrm{N}}^2}\right)$$

and

$$C_{\text{Phase, MI-c.sym., discrete}} = \frac{1}{2}\log_2(M) - \frac{1}{2M} \sum_{m=0}^{M-1} \iint_{\tilde{w}=-\infty}^{\infty} \log_2 \left[\sum_{n=0}^{M-1} \exp\left(-\frac{|\tilde{w} + \tilde{x}_m - \tilde{x}_n|^2 - |\tilde{w}|^2}{\sigma_{\mathrm{N}}^2}\right) \right] f_{\tilde{W}}(\tilde{w})\, \mathrm{d}\tilde{w}$$

where the integration can be replaced by the expected value with respect to $\tilde{W}$ and

$$C_{\text{Phase, MI-c.sym., discrete}} = \frac{1}{2}\log_2(M) - \frac{1}{2M} \sum_{m=0}^{M-1} \mathrm{E}_{\tilde{W}} \left\{ \log_2 \sum_{n=0}^{M-1} \exp\left(-\frac{|\tilde{w} + \tilde{x}_m - \tilde{x}_n|^2 - |\tilde{w}|^2}{\sigma_{\mathrm{N}}^2}\right) \right\}. \quad (3.22)$$

The capacity can be calculated numerically using Monte Carlo simulations by generating a large number of realizations for $\tilde{W}$ and calculating and averaging C. The variables $\tilde{x}_m$ and $\tilde{x}_n$ are discrete and uniformly distributed, and one can sum over all its M values.

3.4.6. Exact Capacity Using Input/Output Complex Symbols

We can generalize the capacity derivation of Section 3.4.5 from a discrete set of input symbols to a continuous input distribution with $\tilde{X} = r_{\mathrm{H}} e^{j\Phi}$ as defined in Section 3.4.1

with a complex channel input $\tilde{x}$ and

$$f_{\tilde{X}}(\tilde{x}_n) = f_{X_a,X_b}(a,b) = \frac{1}{2\pi r_\mathrm{H}}\delta\Big(\sqrt{a^2+b^2} - r_\mathrm{H}\Big)$$

as derived in (3.7). The complex output sample $\tilde{y}$ has the conditional PDF

$$f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{|\tilde{y}-\tilde{x}|^2}{2\sigma^2}\right)$$

with $\sigma^2 = \frac{\sigma_\mathrm{N}^2}{2}$ and $\tilde{x} = r_\mathrm{H}e^{j\varphi}$. The capacity of the continuous-input continuous-output memoryless channel is [64, p. 252]

$$C_{\text{Phase, MI-c.sym.}} = \frac{1}{2}\sup_{f_{\tilde{X}}(\tilde{x})} I(\tilde{X},\tilde{Y}) = \frac{1}{2}\sup_{f_{\tilde{X}}(\tilde{x})} \iint_{\tilde{x}_m=-\infty}^{\infty} f_{\tilde{X}}(\tilde{x}_m)\cdot \iint_{\tilde{y}=-\infty}^{\infty} f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_m)\log_2\left[\frac{f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_m)}{\iint_{\tilde{x}_n=-\infty}^{\infty} f_{\tilde{X}}(\tilde{x}_n)f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_n)\,\mathrm{d}\tilde{x}_n}\right]\mathrm{d}\tilde{y}\,\mathrm{d}\tilde{x}_m.$$

Given the desired signal constellation, the capacity is maximum if the channel inputs $\tilde{x}$ are uniformly distributed on the circle with radius r_H (with a density $f_{\tilde{X}}(\tilde{x})$ as described above), and the supremum in the above term can be omitted. We further substitute $\tilde{Y} := \tilde{W} + \tilde{X}_m$ with $\tilde{W} \sim \mathcal{CN}(0,\sigma_\mathrm{N}^2)$, resulting in

$$f_{\tilde{Y}|\tilde{X}}(\tilde{y}|\tilde{x}_m) = f_{\tilde{W}}(\tilde{w}) = \frac{1}{\pi\sigma_\mathrm{N}^2}\exp\left(-\frac{|\tilde{w}|^2}{\sigma_\mathrm{N}^2}\right)$$

and

$$C_{\text{Phase, MI-c.sym.}} = -\frac{1}{2}\iint_{\tilde{x}_m=-\infty}^{\infty} f_{\tilde{X}}(\tilde{x}_m)\cdot \iint_{\tilde{w}=-\infty}^{\infty}\log_2\left[\iint_{\tilde{x}_n=-\infty}^{\infty} f_{\tilde{X}}(\tilde{x}_n)\exp\left(-\frac{|\tilde{w}+\tilde{x}_m-\tilde{x}_n|^2-|\tilde{w}|^2}{\sigma_\mathrm{N}^2}\right)\mathrm{d}\tilde{x}_n\right] f_{\tilde{W}}(\tilde{w})\,\mathrm{d}\tilde{w}\,\mathrm{d}\tilde{x}_m$$

where the integration over $\tilde{w}$ can be replaced by the expected value with respect to $\tilde{W}$ and

$$C_{\text{Phase, MI-c.sym.}} = -\frac{1}{2}\iint_{\tilde{x}_m=-\infty}^{\infty} f_{\tilde{X}}(\tilde{x}_m)\cdot \mathrm{E}_{\tilde{W}}\left\{\log_2\left[\iint_{\tilde{x}_n=-\infty}^{\infty} f_{\tilde{X}}(\tilde{x}_n)\exp\left(-\frac{|\tilde{w}+\tilde{x}_m-\tilde{x}_n|^2-|\tilde{w}|^2}{\sigma_\mathrm{N}^2}\right)\mathrm{d}\tilde{x}_n\right]\right\}\mathrm{d}\tilde{x}_m$$

and the same for $\tilde{x}_n$

$$C_{\text{Phase, MI-c.sym.}} = -\frac{1}{2} \iint_{\tilde{x}_m=-\infty}^{\infty} f_{\tilde{X}}(\tilde{x}_m) \cdot \mathrm{E}_{\tilde{W}} \left\{ \log_2 \left[\mathrm{E}_{\tilde{x}_n} \left\{ \exp\left(-\frac{|\tilde{w} + \tilde{x}_m - \tilde{x}_n|^2 - |\tilde{w}|^2}{\sigma_{\mathrm{N}}^2} \right) \right\} \right] \right\} \mathrm{d}\tilde{x}_m$$

and the same for $\tilde{x}_m$

$$C_{\text{Phase, MI-c.sym.}} = -\frac{1}{2} \mathrm{E}_{\tilde{x}_m} \left\{ \mathrm{E}_{\tilde{W}} \left\{ \log_2 \left[\mathrm{E}_{\tilde{x}_n} \left\{ \exp\left(-\frac{|\tilde{w} + \tilde{x}_m - \tilde{x}_n|^2 - |\tilde{w}|^2}{\sigma_{\mathrm{N}}^2} \right) \right\} \right] \right\} \right\}. \tag{3.23}$$

The capacity C can be calculated numerically using Monte Carlo simulations by generating a large number of realizations for $\tilde{x}_m$, $\tilde{w}$ and $\tilde{x}_n$, and calculating and averaging C, or by numerical integration. To simplify the numerical integration, given that $f_{\tilde{X}}(\tilde{x})$ is zero for all $\tilde{x}$ in the complex plane except on a circle, we can replace the two double integrals over $\tilde{x}$ by single integrals along the circles, resulting in

$$C_{\text{Phase, MI-c.sym.}} = -\frac{1}{4\pi r_{\mathrm{H}}} \int_{\varphi_m=0}^{2\pi} \iint_{\tilde{w}=-\infty}^{\infty} f_{\tilde{W}}(\tilde{w}) \cdot \log_2 \left[\int_{\varphi_n=0}^{2\pi} \frac{1}{2\pi r_{\mathrm{H}}} \exp\left(-\frac{|\tilde{w} + r_{\mathrm{H}} e^{j\varphi_m} - r_{\mathrm{H}} e^{j\varphi_n}|^2 - |\tilde{w}|^2}{\sigma_{\mathrm{N}}^2} \right) \mathrm{d}\varphi_n \right] \mathrm{d}\tilde{w}\, \mathrm{d}\varphi_m.$$

For numerical integration it is further possible to replace the integrals over φ by summations, i.e.

$$\int_{\varphi=0}^{2\pi} f(\varphi)\, \mathrm{d}\varphi \approx \sum_{i=0}^{S-1} f(i\varphi_\Delta)\, \varphi_\Delta \quad \text{with } \varphi_\Delta = \frac{2\pi}{S} \text{ and, e.g., } S = 256.$$

3.4.7. Comparison of Derived Phase Modulation Capacities

The derived phase modulation capacities of Section 3.4.2 to Section 3.4.6 are shown for different channel SNR $\frac{\sigma_{\mathrm{H}}^2}{\sigma_{\mathrm{N}}^2}$ in Figure 3.6. The differences of the capacities relative to $C_{\text{Phase, MI-c.sym.}}$ of (3.23) are shown in Figure 3.7. The reason for $C_{\text{Phase, MI-c.sym.}}$ being slightly larger than $C_{\text{Phase, MI-angle}}$ of (3.21) is that in the former case the detector has access to both the amplitude and the phase angle of the received symbol, whereas in the later case only the phase angle is known. The figures also include the general capacity of the power-constrained AWGN channel given by the Shannon–Hartley theorem [64]

$$C_{\text{Shannon}} = \frac{1}{2} \log_2 \left(1 + \frac{\sigma_{\mathrm{H}}^2}{\sigma_{\mathrm{N}}^2} \right).$$

It is the upper bound on the information rate when ignoring any watermarking-induced constraint.

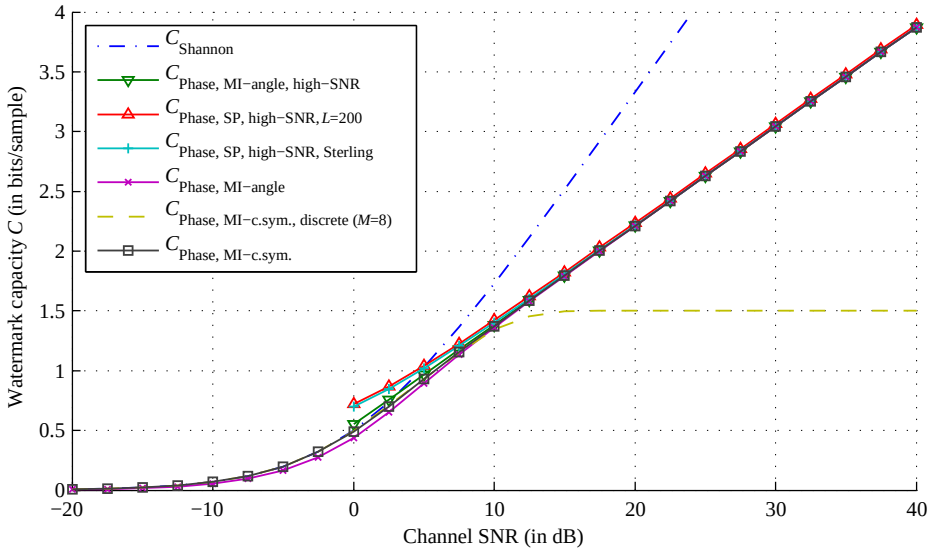


Figure 3.6.: Comparison of the derived phase modulation watermark capacity expressions at different channel SNR $\frac{\sigma_H^2}{\sigma_N^2}$ (see also Figure 3.7).

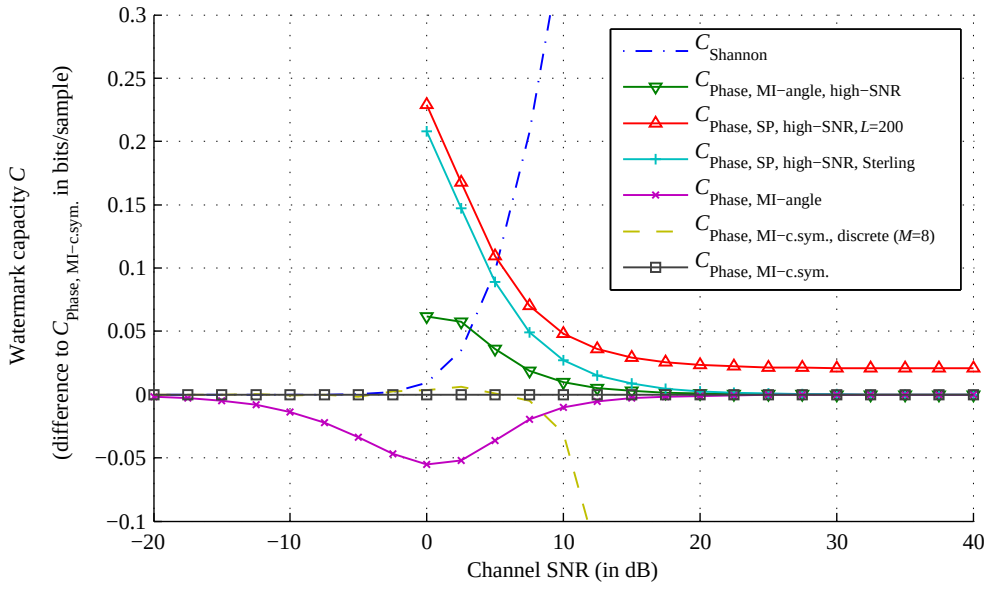


Figure 3.7.: Watermark capacity difference between the derived expressions and $C_{\text{Phase, MI-c.sym.}}$.

3.4.8. Phase Watermarking in Voiced Speech

Given the large capacity of phase modulation based watermarking in unvoiced speech, it would be attractive to extend this approach also to voiced speech. In voiced speech, however, perceptually transparent phase manipulations are much more difficult to achieve, since the signal's phase is indeed perceivable [49, 50, 69, 70]. It is important to maintain 1) the phase continuity of the harmonic components across the analysis/synthesis frames, and 2) the spectral envelope of the signal, including the spectral peaks of the harmonics and the deep valleys in-between them, independent of the position of the analysis window.

We explored possibilities to manipulate or randomize the signal phase in speech while maintaining the two aforementioned perceptual requirements for voiced speech. It follows a brief outline of the performed experiments, which will show that even under idealistic assumptions it is not possible to fully randomize the phase of voiced speech without severely degrading perceptual quality or resorting to perceptual masking.

3.4.8.1. Phase Modulation with the Short-Time Fourier Transform (STFT)

In this experiment, a fully periodic synthetic speech signal with constant pitch of 146 Hz was transformed into DFT domain using a pitch-synchronous short-time Fourier transform (STFT) with a 'square root Hann' analysis window with 50 % overlap and a window length that is an even multiple of the pitch cycle length. Using the same window as synthesis window, the sum of the inverse DFT transforms of the individual frames ('overlap/add synthesis') results in perfect reconstruction of the original signal [71]. To evaluate the suitability of the signal's phase for watermarking, we randomized the phase angle of the complex DFT/STFT coefficients and examined the perceptual degradation as a function of the window length.

Using a short analysis/synthesis window with a window length of 137 ms or 20 pitch cycles, the phase randomization leads to significant perceptual distortion and makes the speech signal sound like produced in a 'rain barrel'. Figure 3.8 shows the long-term high-frequency-resolution DFT spectrum (using a measurement window length of 1 s) of the original signal and the phase-randomized signal created using the 137 ms window. It is apparent that the phase randomization leads to a widening of the spectral peaks due to the limited frequency resolution of the analysis/synthesis window. For example, a pure sinusoidal signal is typically represented by several DFT bins due to the spectral shape of the analysis window. If the phase relation between the different DFT bins is not maintained, the inverse transform is not a pure sinusoid anymore but a mixture of several sinusoids with adjacent frequencies and arbitrary phase relations.

The perceptual degradation decreases when using much longer analysis/synthesis windows in the order of magnitude of 200 pitch cycles. Using a long window corresponds to changing the phase less frequently in time, and also results in an increased frequency resolution. However, the window length determines the time/frequency resolution trade-off of the STFT. For real speech signals such long windows lead to unacceptable smearing in the time domain, and also the pitch would not be constant

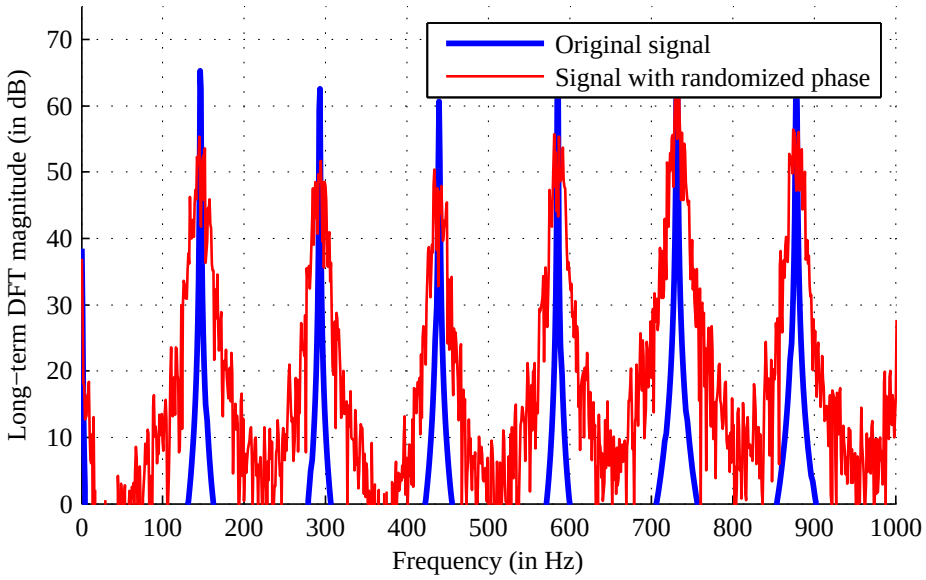


Figure 3.8.: Long-term high-frequency-resolution DFT spectrum with a measurement window length of 1000 ms of 1) a constant pitch periodic speech signal and 2) the same signal with randomized STFT-domain phase coefficients (using an analysis/synthesis window length of 137 ms).

for such long periods.

The 'rain barrel effect' can be eliminated by modulating only those coefficients that are perceptually less important. To verify this, we calculated for every frame of the STFT representation (using a window length of 27 ms or approximately four pitch cycles, and 50 % overlap) the frequency masking curve using the van-de-Par masking model [54] (see also Section 3.3). Then, we randomized only the phase of those DFT coefficients whose magnitude is a certain threshold below the masking curve. We used as test signal a male speech sustained vowel 'a' at a sampling frequency of 16 kHz and a listening level of 82.5 dB_{SPL}. A threshold of 0 dB (randomizing all coefficients below the masking curve) resulted in 83 % of the coefficients being randomized and a certain roughness in the sound, which again results from a widening of spectral peaks (measured over a longer window), namely those peaks where a part of the initial peak is masked. With a threshold of 6 dB (randomizing all coefficients 6 dB below the masking curve, shown in Figure 3.9) 64 % of the phase coefficients were randomized, and the previous roughness in the sound was barely noticeable anymore. However, if a DFT coefficient is masked anyhow, then there is no need to maintain the magnitude of the coefficient and one can proceed with the general masking-based approach discussed in Section 3.3.

Independent of window length and masking, with using a 50% overlap to avoid discontinuities at the frame boundaries, the speech signal is oversampled by a factor of two, and the STFT representation contains twice as many coefficients as the time domain signal. As a consequence for watermarking, the frequency coefficients are not

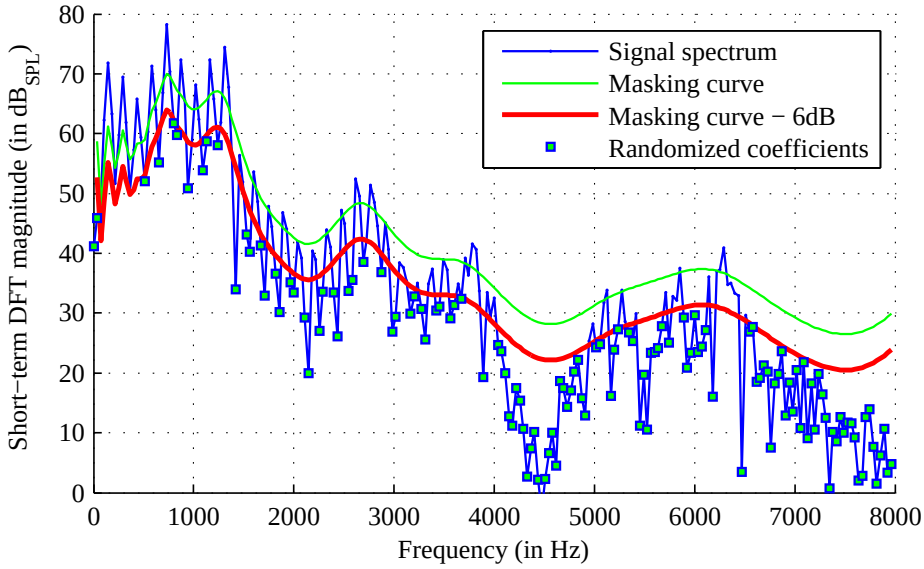


Figure 3.9.: Randomization of the phase of those coefficients DFT that are masked and have a magnitude of less than 6 dB below the masking curve.

independent, and it is not possible to independently randomize all phase coefficients and re-obtain the same coefficients after resynthesis and analysis. An approach to solve this problem is the use of a critically sampled orthogonal transform instead of the STFT. This is discussed in the following subsection.

3.4.8.2. Phase Modulation with Lapped Orthogonal Transforms

The use of a critically sampled orthogonal transform for analysis and synthesis allows to obtain a perfect reconstruction of the embedded watermark data. In the following, we discuss the use of extended and modulated lapped transforms (ELT and MLT, [72, 73, 74, 75]) for watermarking by phase modulation.

Extended Lapped Transform (ELT) The ELT is a critically sampled orthogonal lapped transform with an arbitrary window length L [73, 74]. Just as the more popular MLT, the ELT is a perfect reconstruction cosine-modulated filterbank defined by the same basis functions, but generalized from a fixed window length $L = 2M$ to a variable window length that is typically larger than $2M$, where M denotes the transform size or number of filter channels.

The idea for watermarking is to use an ELT with few coefficients (filter channels) and long windows, and to modify again the signal's phase in the frequency domain. As discussed in Section 3.4.1.1, we denote with modifying the MLT or ELT phase the replacement of the original transform domain subband signals by watermark subband signals with identical short-term power envelopes, since this preserves the spectral envelope of the host signal. The orthogonality of the ELT and MLT assures

that the embedded information is recoverable. A low number of coefficients (which equals the update interval or the hop size) and long windows implicate that at any given time there is a large number of overlapping windows. The hypothesis for this experiment is that if each window has the desired magnitude spectrum, then also the sum of the overlapping windows has the desired magnitude spectrum, no matter of the position the spectrum being measured at. As such, it should be possible to recreate the magnitude spectrum of voiced speech while continuously manipulating the phase spectrum. We evaluate this approach with a small experiment.

We transformed a synthetic stationary constant pitch speech signal into the frequency domain using an ELT as described in [74] with M filter channels and with a prototype filter (window shape) with stopband attenuation A_s, designed using the Kaiser window approach of [76]. The window length L is a result of M and A_s. Since there is no direct notion of phase in the frequency domain of the ELT—it is a real-valued transform—we measured the variance in each filter channel and replaced the original signal in each frequency channel by Gaussian noise with equal variance. The modified signal was then transformed back to time domain.

There are fundamentally contradicting requirements on the parameter M. On the one hand, the number of filter channels determines the frequency resolution of the filter bank, and in order to accurately maintain the sharp spectral peaks in voiced speech, M must be large. For example, to obtain approximately ten subbands in-between two harmonics of a speech signal with a pitch of 150 Hz, one must set $M = 256$ at a sample rate of $f_s = 8\,\mathrm{kHz}$ (resulting in a subband width of $\frac{f_s}{2M} \approx 15.6\,\mathrm{Hz}$). With $A_\mathrm{s} = 50\,\mathrm{dB}$ this results in a window length of 768 ms ($L = 6144$). But even with a frequency resolution as high as this, and even though there is an overlap of 24 different windows at any given time, there is still a significant widening of the spectral peaks (see Figure 3.10). This widening shows itself with the same 'rain barrel effect' as when using the STFT. Additionally, such a long window length results in unacceptable smearing in the time domain. Decreasing M, for example to $M = 64$ and a window length of 112 ms ($A_\mathrm{s} = 30\,\mathrm{dB}$), makes the speech signal more similar to colored noise.

Just as in the STFT case, the 'rain barrel effect' at large M can be eliminated by not randomizing the perceptually most important subchannels. This was confirmed with a small experiment that kept 10 % of the $M = 256$ coefficients unmodified, namely those with the largest power.

Modulated Lapped Transform (MLT) Reducing the window length L of the ELT to twice the transform size M, $L = 2M$, and using a sine window results in the modified lapped transform (MLT) [74, 75]. Applying the same phase randomization as with the ELT, the results are very similar. For identical M, the randomized MLT sounds slightly more noisy than the randomized ELT. For large M, the perceptual results become almost identical for equal effective window lengths (considering at the ELT window only its time-domain 'main lobe').

At a sampling frequency of 8 kHz, an MLT with $M = 64$ has a window length of 16 ms, which is a commonly used window length for speech signal coding. Using this window length and again randomizing the coefficients results in a very noise-like

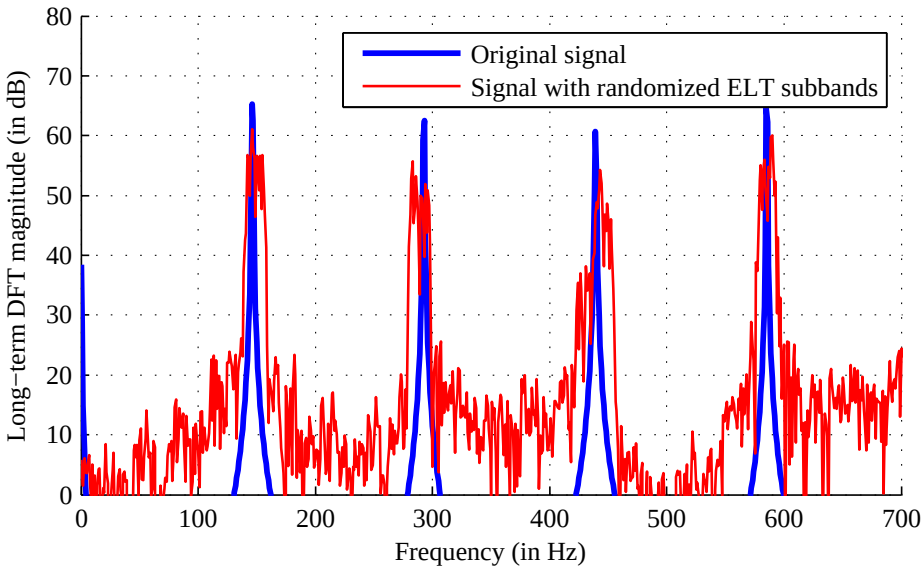


Figure 3.10.: Constant pitch speech signal with original and randomized ELT coefficients using 256 filter channels and a 768 ms-window. The randomization leads to a widening of the spectral peaks. (The plotted DFT spectrum is measured using a 1000 ms-window.)

speech signal. This effect can again be mitigated by randomizing only the perceptually less important coefficients. However, even when keeping 50 % of the coefficients unmodified (again those with the largest power), a perceptual distortion is still audible. Because of the short window length, the distortion is audible as noise. It is expected that the use of a masking model instead of the power criterion to select the relevant coefficients would decrease the audibility of the distortion.

3.4.8.3. Conclusion

We conclude that in voiced speech, in contrast to unvoiced speech, it is not possible to fully randomize the signal's phase without 1) severely degrading perceptual quality or 2) resorting to perceptual masking. Time-frequency transforms allow direct access to the signal's phase, but in order to avoid discontinuities at frame boundaries, block transforms can only be used with overlapping windows. While the STFT has the most intuitive notion of phase, it is an oversampled representation whose phase coefficients are not independent and, thus, difficult to embed and re-extract. This problem can be solved with ELTs and MLTs, but they are subject to the same fundamental trade-off between time and frequency resolution as the STFT. Perceptual artifacts can be eliminated by restricting the phase modulations to masked signal components. However, this is not a useful approach for watermarking, since masked components can be replaced altogether instead of only their phase being modulated.

Note that phase modifications are possible if performed at a much lower rate than

presented herein, for example by randomizing only the initial phase of each voiced segment. Several such methods were previously proposed but are limited in the achievable data rate [14, 47, 48, 45, 46].

3.5. Experimental Comparison

In this section, we compare the derived watermark capacities of the ideal Costa scheme (ICS), watermarking based on frequency masking, and watermarking based on phase modulation.

3.5.1. Experimental Settings

The experimental settings for each method were chosen as follows.

Ideal Costa Scheme The capacity in (3.1) was evaluated using two different scenarios, 1) assuming a fixed mean squared error distortion,

$$C_{\text{ICS, MSE=-25dB}}: \qquad \left.\frac{\sigma_{\text{W}}^2}{\sigma_{\text{H}}^2}\right|_{\text{dB}} = -25\,\text{dB},$$

and 2) assuming a fixed watermark to channel noise ratio (WCNR),

$$C_{\text{ICS, WCNR=-3dB}}: \qquad \left.\frac{\sigma_{\text{W}}^2}{\sigma_{\text{N}}^2}\right|_{\text{dB}} = -3\,\text{dB}.$$

While the first case corresponds to a fixed perceptual distortion of the clean watermarked signal (before channel), the second case corresponds to a fixed perceptual distortion of the noise-corrupted watermarked signal (after channel, assuming that the watermark signal is masked by the channel noise).

Frequency Masking For watermarking based on frequency masking, the watermark capacity C_{Mask} of (3.3) is used, with the watermark power $\sigma_{\text{W,mask}}^2$ being $\kappa = -12\,\text{dB}$ below the masking threshold of the masking model.

Phase Modulation In principle, the capacity of watermarking by phase modulation is given by (3.23) and can be expressed as a function $C_{\text{Phase}}(\sigma_{\text{H}}^2, \sigma_{\text{N}}^2)$, with σ_{H}^2 being equivalent to r_{H}^2. However, two additional factors need to be taken into consideration when dealing with real-world speech signals: First, as we showed in Section 3.4.8, the phase modulation is possible in unvoiced speech only, and second, the energy in a speech signal is unevenly distributed between voiced and unvoiced regions. For the experiment, it was assumed that a fraction γ of the speech signal is unvoiced speech, the average power (or variance) of which is a multiplicative factor ρ above or below the average speech signal power σ_{H}^2. The watermark capacity then evaluates to

$$C_{\text{Phase, speech}} = \gamma \cdot C_{\text{Phase}}(\rho\sigma_{\text{H}}^2, \sigma_{\text{N}}^2). \tag{3.24}$$

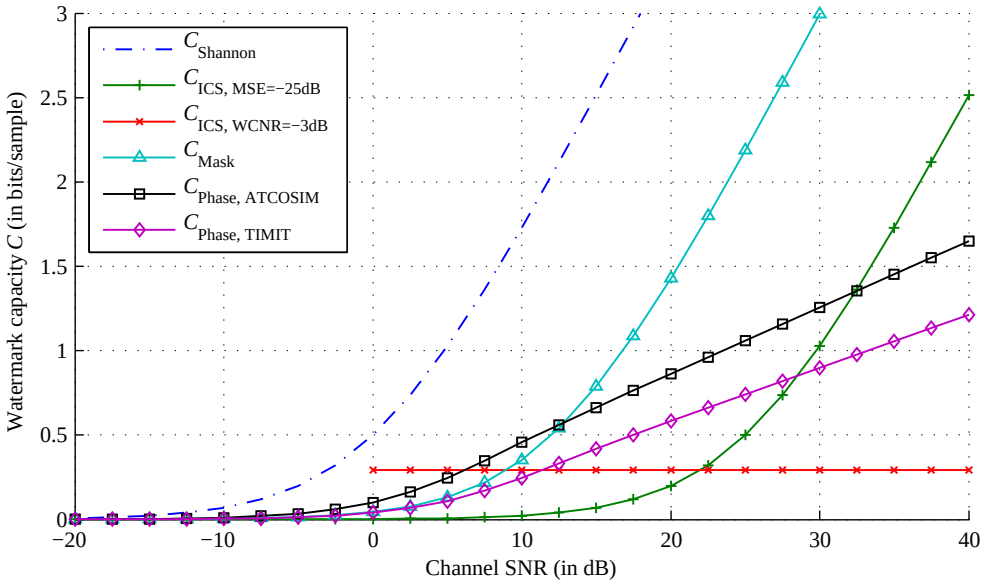


Figure 3.11.: Watermark capacity for quantization, masking and phase modulation based speech watermarking. The abbreviations are explained in Section 3.5.1.

We determined the parameters γ and ρ with a small experiment, again using ten randomly selected utterances (one per speaker) of the ATCOSIM corpus (see Chapter 7). After stripping silent regions (defined by a 43ms-Kaiser-window-filtered intensity curve being more than 30 dB below its maximum for longer than 100ms), each utterance was individually normalized to unit variance, i.e., $\sigma_{\text{H}}^2 = 1$. Then, a voiced/unvoiced segmentation was performed using PRAAT [77], and the total length and average power measured separately for all voiced and all unvoiced segments. The result is a fraction of $\gamma = 47\,\%$ being unvoiced, and the average power in unvoiced being $-4.6\,\text{dB}$ below the overall signal power ($\rho = 0.35$). As a consequence, the unvoiced segments contain 16 % of the overall signal energy. Performing the same analysis on the 192 sentences of the TIMIT Core Test Set [78] results in $\gamma = 38\,\%$, $\rho = 0.16$ ($-8\,\text{dB}$), and only 6 % of the total energy located in unvoiced segments. This difference can be attributed to the slower speaking style in TIMIT compared to the ATCOSIM corpus. The resulting watermark capacities based on the two databases and using (3.24) are denoted $C_{\text{Phase,ATCOSIM}}$ and $C_{\text{Phase,TIMIT}}$.

3.5.2. Results

The watermark capacities C_{ICS}, C_{Mask} and C_{Phase} for quantization, masking, and phase modulation based speech watermarking are shown in Figure 3.11. As in Figure 3.6, the plot includes for comparison the Shannon capacity C_{Shannon} of the AWGN channel.

3.6. Conclusions

Watermarking in perceptually relevant speech signal components, as it is required in certain applications that need robustness against lossy coding, is best performed using variants of the ideal Costa scheme, and its capacity C_{ICS} is shown in Figure 3.11. In contrast, in the application of interest in this work—speech watermarking for analog legacy system enhancement—watermarking in perceptually irrelevant speech signal components is possible. Then, perceptual masking and the auditory system's insensitivity to phase in unvoiced speech can be exploited to increase the watermark capacity (see C_{Mask} and C_{Phase} in Figure 3.11).

Using auditory masking leads to a significant capacity gain compared to the ideal Costa scheme. For high channel SNRs the masking-based capacity is larger than all other watermark capacities in consideration.

The watermark capacity in the phase of unvoiced speech can be derived using the sphere packing analogy or using the related concept of mutual information. In unvoiced speech and for high channel SNRs, the capacity is roughly half of the Shannon capacity plus half a bit per independent sample. The overall watermark capacity is signal-dependent. For fast-paced air traffic control speech at low and medium channel SNR, which is the application of interest in this work, the phase modulation approach outperforms the masking-based approach in terms of watermark capacity.

The remainder of this thesis focuses on the phase modulation approach. Some form of auditory masking is used in many, if not even most, state-of-the art audio watermarking algorithms, and is a well-explored topic [13]. In contrast, a complete randomization of the speech signal's phase has, to our best knowledge, not been previously proposed in the context of watermarking. It offers an opportunity for a novel contribution to the area of speech watermarking and opens a window to implementations that could possibly outperform current state-of-the-art methods.

Note that the phase modulation approach and the conventional masking-based approach do not contradict each other in any significant way. In fact, it is expected that a combination of the two approaches will lead to a further increase in watermark capacity.

Chapter 4

Watermarking Non-Voiced Speech

We present a blind speech watermarking algorithm that embeds the watermark data in the phase of non-voiced speech by replacing the excitation signal of an autoregressive speech signal representation. The watermark signal is embedded in a frequency subband, which facilitates robustness against bandpass filtering channels. We derive several sets of pulse shapes that prevent intersymbol interference and that allow the creation of the passband watermark signal by simple filtering. A marker-based synchronization scheme robustly detects the location of the embedded watermark data without the occurrence of insertions or deletions.

In light of the potential application to analog aeronautical voice radio communication, we present experimental results for embedding a watermark in narrowband speech at a bit rate of 450 bit/s. The adaptive equalization-based watermark detector not only compensates for the vocal tract filtering, but also recovers the watermark data in the presence of non-linear phase and bandpass filtering, amplitude modulation and additive noise, making the watermarking scheme highly robust.

Parts of this chapter have been published in K. Hofbauer, G. Kubin, and W. B. Kleijn, "Speech watermarking for analog flat-fading bandpass channels," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, revised and resubmitted.

As shown in Chapter 3, the combination of established watermarking theory with a well-known principle of speech perception leads to a substantial improvement in theoretical capacity. Based on this finding, the following chapter develops a practical watermarking scheme that is aimed at analog legacy system enhancement. Although the method is of general value, many of our design choices as well as the choice of attacks that we consider are motivated by the application of watermarking to the aeronautical voice radio communication between aircraft pilots and air traffic control (ATC) operators as described in detail in Chapter 6.

In the remainder of this chapter, Section 4.1 proposes the watermarking scheme, also addressing many practical issues such as synchronization and channel equalization. After a discussion of certain implementation aspects in Section 4.2, experimental results are presented in Section 4.3. Finally, we discuss our results and draw conclusions in Section 4.4 and 4.5.

4.1. Theory

In Chapter 3 we have shown the large theoretical capacity of watermarking by replacing the phase of the host signal. Motivated by this finding, this section presents a practical speech watermarking scheme that is based on this principle. The new method also addresses the two major difficulties with the theoretical approach: Firstly, the human ear is only partially insensitive to phase modifications in speech, and secondly, the transmission channel depends on the application and is in most cases not only a simple AWGN channel. We assume in the following all signals to be discrete-time signals with a sample rate f_s.

4.1.1. Speech Signal Model

Our method is based on an autoregressive (AR) speech signal model. This is a commonplace speech signal representation, which models the resonances of the vocal tract and is widely used in speech coding, speech synthesis, and speech recognition (cf. [58, 60, 79]). Consequently, we consider certain temporal sections of a speech signal $s(n)$ as an outcome of an order P autoregressive signal model with time-varying predictor coefficients $\mathbf{c}(n) = [c_1(n), \ldots, c_P(n)]^{\mathrm{T}}$ and with a white Gaussian excitation $e(n)$ with time-variant gain $g(n)$, such that

$$s(n) = \sum_{m=1}^{P} c_m(n) s(n-m) + g(n)e(n).$$

4.1.2. Transmission Channel Model

We focus on analog legacy transmission channels such as the aeronautical radio or the PSTN telephony channel. As an approximation, we assume the channel to be an analog AWGN filtering channel with a limited passband width. The transmission channel attacks are listed in the left column of Table 4.1.

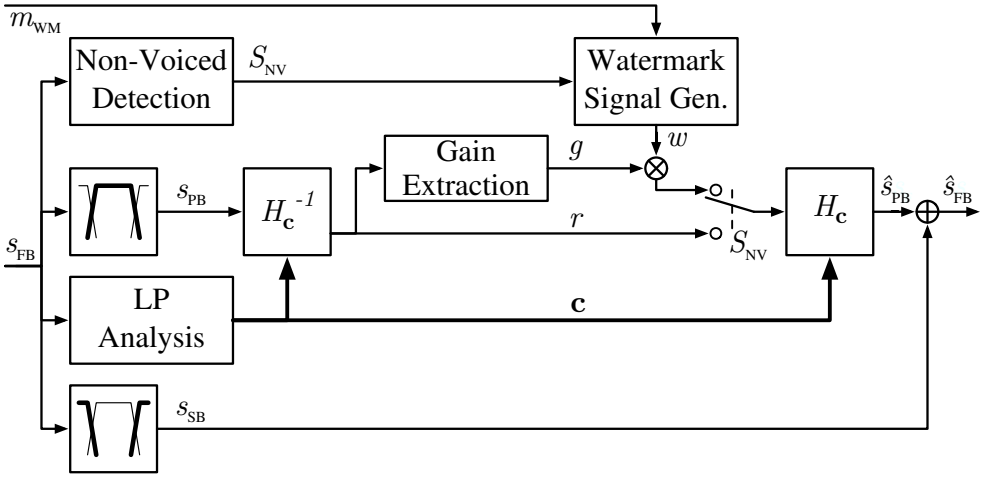


Figure 4.1.: Watermark embedding: Non-voiced speech (or a subband thereof) is exchanged by a watermark data signal.

4.1.3. Watermark Embedding Concept

The presented speech signal model facilitates the embedding of a data signal by replacing the AR model phase of the host speech signal (see Section 3.4.1.1). An important property of the signal model is that there is no perceptual difference between different realizations of the Gaussian excitation signal $e(n)$ for non-voiced speech [42]. It is possible to exchange the white Gaussian excitation signal $e(n)$ by a white Gaussian *data* signal $\hat{e}(n)$ that carries the watermark information. The signal thus forms a hidden data channel within the speech signal. We denote with 'non-voiced speech' all speech that is not voiced, comprising unvoiced speech and pauses.

The hidden data channel is restricted both in the temporal and in the spectral domain. In the temporal domain, the replacement of $e(n)$ by $\hat{e}(n)$ can take place only when the speech signal is not voiced, since the model as defined in Section 4.1.1 is accurate only for non-voiced speech. The speech parts that are voiced and should remain unmodified are detected based on their acoustic periodicity using an autocorrelation method [80]. In the spectral domain, only a subband bandpass component $e_{PB}(n)$ of $e(n)$ can be replaced by the data signal $\hat{e}(n)$. This accommodates the bandpass characteristic of the transmission channel, and the embedding band width and position must be selected such that they fully lie within the passband of the transmission channel. The stopband component $e_{SB}(n) = e(n) - e_{PB}(n)$ must be kept unchanged, since in many practical applications the exact channel bandwidth is not known a priori. This is realized by adding an unmodified secondary path for the stopband component $s_{SB}(n)$.

Figure 4.1 shows an overview of the embedding scheme. First, the voiced and the non-voiced time segments of the fullband input speech signal $s_{FB}(n)$ are identified and indicated by a status signal S_{NV}. Then, the predictor coefficients $\mathbf{c}(n)$ are calculated and regularly updated using linear prediction (LP) analysis. The signal $s_{FB}(n)$ is decomposed into two subbands (without downsampling), and a LP error filter ($H_{\mathbf{c}}^{-1}$)

Table 4.1.: Transmission Channel and Hidden Data Channel Model

Transmission Channel Attacks	*System-Inherent Attacks*
DA/AD conversion (resampling)	Random channel availability S_{NV}
Desynchronization	Time-variant gain g
AWGN	Time-variant all-pole filtering $H_{\mathbf{c}}$
Bandpass filtering	Additive stopband signal s_{SB}
Magnitude and phase distortion	

with predictor coefficients $\mathbf{c}(n)$ is applied on the passband component $s_{\mathrm{PB}}(n)$, resulting in the passband error signal $r(n) = g(n)e(n)$. The LP analysis is based on $s_{\mathrm{FB}}(n)$, since otherwise the LP error filter would try to compensate the bandpass characteristic of $s_{\mathrm{PB}}(n)$. In voiced speech segments, the passband signal is resynthesized using the original error signal $r(n)$ and the LP synthesis filter $H_{\mathbf{c}}$. In contrast, the non-voiced speech segments (including unvoiced speech and inactive speech) are resynthesized from a spectrally shaped watermark data signal $w(n)$, on which the original gain $g(n)$ of the error signal and the LP synthesis filter $H_{\mathbf{c}}$ is applied. The unmodified stopband component $s_{\mathrm{SB}}(n)$ is added to the watermarked passband signal $\hat{s}_{\mathrm{PB}}(n)$, resulting in the watermarked speech $\hat{s}_{\mathrm{FB}}(n)$.

4.1.4. Hidden Data Channel Model

The watermark detector is ultimately interested in the payload watermark m_{WM}. However, the watermark channel carrying the hidden data signal is subject to a number of channel attacks, which are listed in Table 4.1. They are either inherent to the embedding procedure of Section 4.1.3, or induced by the transmission channel. While the system-inherent attacks are deterministic and known to the watermark embedder, they are randomly varying quantities and unknown to the watermark detector.

The following sections further refine the embedding concept of Section 4.1.3 to counteract the channel attacks defined in Table 4.1. In principle, the payload watermark m_{WM} must be transformed into a watermark signal $w(n)$ using a modulation scheme that is either robust against the channel attack or enables the detector to estimate and revert the attack.

4.1.5. Watermark Signal Generation

In the following, we describe the composition of the real-valued watermark signal $w(n)$. It is subject to conditions that result from the perceptual transparency, data transmission rate and robustness requirements.

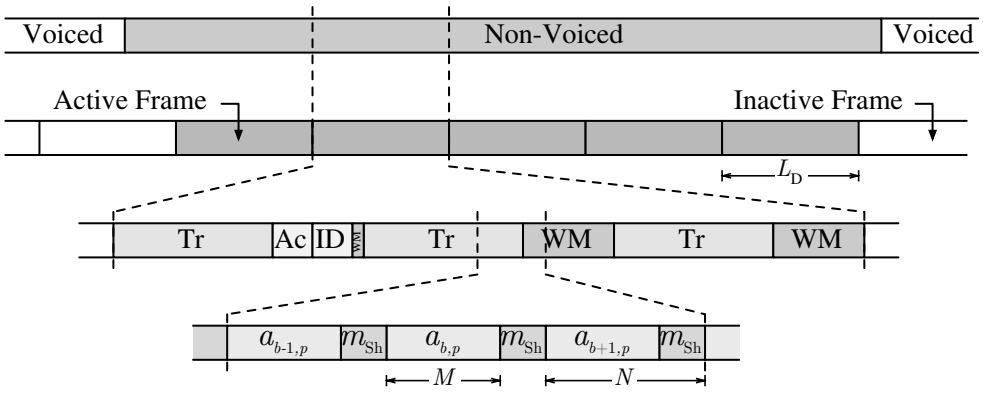


Figure 4.2.: Frame and block structure including watermark (WM), training (Tr), active frame (Ac) and frame identification (ID) regions, which carry the data symbols $a_{b,p}$ interleaved with the spectral shaping symbols m_{Sh}.

4.1.5.1. Bit Stream Generation

The encoded payload data is interleaved with additional auxiliary symbols in order to facilitate the subsequent processing steps. As will be discussed in Section 5.1.3, we use a fixed frame grid with frames of equal length L_{D}. Each frame that lies entirely within a non-voiced segment of the input signal is considered as *active* frame and consists of L_{WM} watermark symbols m_{WM} representing encoded payload data, L_{Tr} training symbols m_{Tr} for signal equalization in the receiver (Section 4.1.7), L_{Ac} preamble symbols m_{Ac} for marking active frames (Chapter 5), L_{ID} preamble symbols m_{ID} for sequentially numbering active frames with a consecutive counter (Chapter 5), and L_{Sh} spectral shaping symbols m_{Sh} for pulse shaping and fulfilling the bandpass constraint (Section 4.1.5.2). The symbols are real-valued, and interleaved with each other as shown in Figure 4.2. While m_{WM}, m_{Tr}, m_{Ac} and m_{ID} could in principle be multi-level and multi-dimensional symbols, trading data rate against robustness, we use in the implementation of Section 4.2 single-level bipolar one-dimensional symbols, and denote with A the sequence of these interleaved data symbols and with $a_{b,p}$ the individual terms of A as defined in the next section. The spectral shaping symbols m_{Sh} are multi-level real scalars and not part of the data sequence A.

When using the signal model presented in Section 4.1.1, the set of pseudo-random data, training and preamble symbols must consist of independent symbols, and have an amplitude probability distribution that is zero mean, unit variance, and Gaussian. However, we previously showed that violating the requirement of Gaussianity does not severely degrade the perceptual quality [3].

4.1.5.2. Pulse Shaping

The aim of pulse shaping is to transform the data sequence A into a sequence of samples $w(n)$ that

1. has the same predefined and uniform sampling frequency f_s as the processed speech signal,
2. has no intersymbol interference (ISI) between the data symbols A,
3. has a defined boxcar (bandpass) power spectral density with lower and upper cut-off frequencies f_L and f_H that lie within the passband of the transmission channel and with $f_\mathrm{H} \leq \frac{f_s}{2}$, and
4. carries as many data symbols as possible per time interval (i.e., the symbol embedding rate f_a is maximum).

With a simple one-to-one mapping from A to $w(n)$ the signal $w(n)$ does not have the desired spectral shape and the data is not reliably detectable since the Nyquist rate, which is twice the available transmission channel bandwidth, is exceeded. To solve this problem, we interleave the signal $w(n)$ with additional samples m_Sh to reduce the rate f_a at which the data symbols A are transmitted to below the Nyquist rate. Additionally, the values of the samples m_Sh are set such that the resulting signal $w(n)$ has the desired spectral shape.

Finding the correct quantities, distributions and values for the spectral shaping samples m_Sh given specified cut-off and sampling frequencies is a non-trivial task. However, transforming a data sequence into a band-limited signal is in many aspects the inverse problem to representing a continuous-time band-limited signal by a sequence of independent and possibly non-uniformly spaced samples, and the theories of non-uniform sampling and sampling and interpolation of band-limited signals can be applied to this problem (cf. [81, 82, 83]). We consider the desired watermark signal $w(t)$ as the band-limited signal, the given data symbol sequence A with an average symbol rate f_a as the non-uniformly spaced samples of this signal, and the required pulse shape $y(t)$ as the interpolation function to reconstruct the band-limited signal from its samples. The wanted symbols m_Sh are obtained by sampling the reconstructed signal $w(t)$ at defined time instants. The above four requirements on $w(n)$ then translate to

1. constraining the sampling instants onto a fixed grid with spacing $1/f_s$ or a subset thereof,
2. requiring the samples of the signal to be algebraically independent, i.e., any set of samples A defines a unique signal $w(t)$ with the desired properties,
3. requiring the signal to be band-limited to f_L and f_H, and
4. requiring the symbol rate f_a to be close to the Nyquist rate of $2(f_\mathrm{H} - f_\mathrm{L})$.

It is known from theory that these requirements can be fulfilled only for certain combinations of f_s, f_L and f_H (e.g., [83]). The permission of non-uniform sampling loosens the constraints on the permissible frequency constellations.

A common approach for a practical implementation is to choose the parameters such that in a block of N samples on a sampling grid with uniform spacing $T_s = 1/f_s$

the first M samples are considered as independent samples and fully describe the band-limited signal (e.g., [84]). The remaining $N - M$ samples are determined by the bandwidth-constraint and can be reconstructed with suitable interpolation functions. Analogously, the first M samples are the information-carrying data samples, and the remaining $N - M$ samples are the spectral shaping samples, determined by the interpolation function.

To formalize the problem, let A be the sequence of interleaved watermark, training and preamble symbols as defined in Section 4.1.5.1 and a denote the terms of this sequence. A is divided into non-overlapping and consecutive blocks of M terms, and $a_{b,p}$ denotes the p'th term ($p = 1 \dots M$) in the b'th block. The terms $a_{b,p}$ of A are considered as the samples of the band-limited signal $w(t)$, which is non-uniformly sampled at the time instants $\tau_{b,p} = t_p + bNT_s$ with $t_p = (1 - p)T_s$ and $T_s = 1/f_s$. The desired signal $w(t)$ can then be constructed with [85]

$$w(t) = \sum_{b=-\infty}^{\infty} \sum_{p=1}^{M} a_{b,p} y_{b,p}(t), \tag{4.1}$$

where $y_{b,p}(t)$ is the reconstruction pulse shape that is time-shifted to the b'th block and used for the p'th sample in each block, and with subsequent resampling of $w(t)$ at the time instants $\tau = nT_s$.

To have all previously stated requirements fulfilled, N, M, f_s, f_L, f_H and y_p must be chosen accordingly. We provide suitable parameter sets and pulse shapes for different applications scenarios in Section 4.2.2.

4.1.6. Synchronization

We deal with the different aspects of synchronization between the watermark embedder and the watermark detector extensively in Chapter 5. In summary, sampling timing and bit synchronization as well as signal synthesis and analysis synchronization are inherently taken care of by the embedded training sequences and the adaptive equalizer in the watermark detector. Data frame synchronization is achieved by using a fixed frame grid and the embedding of preambles.

4.1.7. Watermark Detection

Figure 4.3 gives an overview on the components of the watermark detector, which follows a classical front-end–equalization–detection design [86].

4.1.7.1. Detector Front End

In the first step the received analog watermarked speech signal is sampled. In the presence of an RLS adaptive equalizer as proposed in the next subsection, the sampling operation does not need to be synchronized to the watermark embedder. The channel delay is estimated and the signal is realigned with the frame grid. These different aspects of synchronization are discussed in Chapter 5. In the remainder of this chapter we assume for the received signal path that perfect synchronization has been achieved.

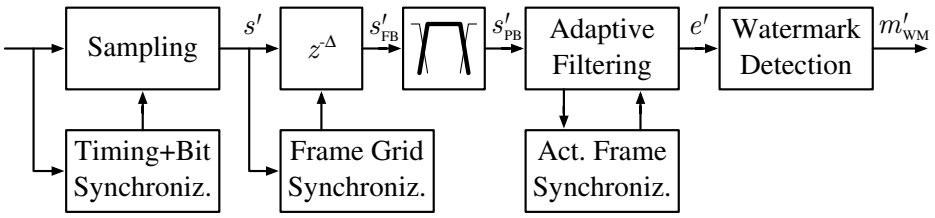


Figure 4.3.: Watermark detector including synchronization, equalization and watermark detection.

From the discrete-time speech signal $s'_{\mathrm{FB}}(n)$ the out-of-band components $s'_{\mathrm{SB}}(n)$ are again removed by an identical bandpass filter as in the embedder, resulting in the passband signal $s'_{\mathrm{PB}}(n)$.

4.1.7.2. Channel Equalization

The received signal is adaptively filtered to compensate for the all-pole filtering in the embedder as well as for the unknown transmission channel filtering. We make use of the training symbols m_{Tr} embedded in the watermark signal $w(n)$ in order to apply a recursive least-squares (RLS) equalization scheme. The RLS method is chosen over other equalization techniques such as LMS because of its rate of convergence and tracking behavior, and the availability of efficient recursive implementations that do not require matrix inversion [87].

We use an exponentially weighted and non-uniformly updated RLS adaptive filter in an equalization configuration, which tracks the joint time-variation of the vocal tract filter and the transmission channel [87]: Let K be the number of taps and λ $(0 < \lambda \leq 1)$ be the exponential weighting factor or forgetting factor of the adaptive filter and α be a fixed constant that depends on the expected signal-to-noise ratio in the input signal. Also let $\mathbf{u}(n) = [s'_{\mathrm{PB}}(n+K-1), \ldots, s'_{\mathrm{PB}}(n)]^{\mathrm{T}}$ be the K-dimensional tap input vector and $\mathbf{w}(n)$ the K-dimensional tap-weight vector at time n, with $\mathbf{w}(0) = \mathbf{0}$. We initialize the inverse correlation matrix with $\mathbf{P}(0) = \delta^{-1}\mathbf{I}$, using the regularization parameter $\delta = \sigma_s^2(1-\lambda)^{\alpha}$ with σ_s^2 being the variance of the input signal $s'_{\mathrm{PB}}(n)$. For each time instant $n = 0, 1, 2, \ldots, \infty$ we iteratively calculate

$$\begin{aligned}
\mathbf{q}(n) &= \mathbf{P}(n)\mathbf{u}(n) \\
\mathbf{k}(n) &= \frac{\mathbf{q}(n)}{\lambda + \mathbf{u}^H(n)\mathbf{q}(n)} \\
\tilde{\mathbf{P}}(n) &= \lambda^{-1}(\mathbf{P}(n) - \mathbf{k}(n)\mathbf{q}^H(n)) \\
\mathbf{P}(n+1) &= \tfrac{1}{2}(\tilde{\mathbf{P}}(n) + \tilde{\mathbf{P}}^H(n))
\end{aligned}$$

where * denotes complex conjugation and ^{H} denotes Hermitian transposition. We introduced the last equation to assure that $\mathbf{P}(n)$ remains Hermitian for numerical stability. The same effect can be achieved computationally more efficient by calculating only the upper or lower triangular part of $\tilde{\mathbf{P}}(n)$ and filling in the rest such that

Hermitian symmetry is preserved. If no training symbol is available, the tap-weights are not updated, and $\mathbf{w}(n+1) = \mathbf{w}(n)$. If a training symbol is available for the current time instant, the error signal e_{R} and the tap weights are updated with

$$e_{\mathrm{R}}(n) = m_{\mathrm{Tr}}(n) - \mathbf{w}^H(n)\mathbf{u}(n)$$
$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mathbf{k}e_{\mathrm{R}}^*(n).$$

The output $e'(n)$ of the RLS adaptive filter is

$$e'(n) = \mathbf{w}^H(n)\mathbf{u}(n).$$

After performing the active frame detection (see Section 5.1.3) an equalizer retraining is performed. The active frame markers m_{Ac} can be used as additional training symbols, which improves the detection of the subsequent frame ID symbols m_{ID} (see Figure 4.2).

4.1.7.3. Detection

An equalization scheme as presented in the previous subsection avoids the necessity of an expensive signal detection scheme, effectively shifting complexity from the detection stage to the equalization stage [86]. Consequently, the embedded watermark data is detected after synchronization and equalization using a simple minimum Euclidean distance metric.

4.2. Implementation

This section presents details of our signal analysis and pulse shaping implementation, which are required to reproduce the experimental results presented in Section 4.3. In this section a number of equations are acausal to simplify notation. For implementation, the relationships can be made causal by the addition of an input signal buffer and appropriate delays.

4.2.1. Signal Analysis

We first address the decomposition of the input signal into the model components. The perfect-reconstruction subband decomposition of the discrete-time speech signal $s_{\mathrm{FB}}(n)$ with sample rate f_s into the embedding-band and out-of-band speech components $s_{\mathrm{PB}}(n)$ and $s_{\mathrm{SB}}(n)$ is obtained with a Hamming window design based linear phase finite impulse response (FIR) bandpass filter of order N_{BP} (even) with filter coefficients $\mathbf{h} = [h_0, \ldots, h_{N_{\mathrm{BP}}}]^{\mathrm{T}}$, resulting in

$$s_{\mathrm{PB}}(n) = \sum_{m=0}^{N_{\mathrm{BP}}} h_m s_{\mathrm{FB}}(n - m + \tfrac{N_{\mathrm{BP}}}{2})$$

and $s_{\mathrm{SB}}(n) = s_{\mathrm{FB}}(n) - s_{\mathrm{PB}}(n)$. The filter bandwidth $W_{\mathrm{PB}} = f_{\mathrm{H}} - f_{\mathrm{L}}$ must be chosen in accordance to the transmission channel and the realizable spectral shaping configuration as described in Section 4.1.5.2. Complete parameter sets for three different transmission channel types are provided in Section 4.2.2.

The linear prediction coefficient vector $\mathbf{c}(n)$ is obtained from solving the Yule-Walker equations using the Levinson-Durbin algorithm, with

$$\mathbf{c}(n) = \mathbf{R}_{\mathbf{ss}}^{-\mathbf{1}}(n)\mathbf{r}_{\mathbf{ss}}(n), \tag{4.2}$$

where $\mathbf{R}_{\mathbf{ss}}(n)$ is the autocorrelation matrix and $\mathbf{r}_{\mathbf{ss}}(n)$ is the autocorrelation vector of the past input samples $s_{\mathrm{FB}}(n-m)$ using the autocorrelation method and an analysis window length of 20 ms, which is common practice in speech processing [79]. We update $\mathbf{c}(n)$ every 2 ms in order to obtain a smooth time-variation of the adaptive filter. If computational complexity is of concern, the update interval can be increased and intermediate values can be obtained using line spectral frequency (LSF) interpolation [58].

The LP error signal $r(n)$ is given by the LP error filter $H_{\mathbf{c}}^{-1}$ with

$$r(n) = s_{\mathrm{PB}}(n) - \sum_{m=1}^{P} c_m(n) s_{\mathrm{PB}}(n-m) = g(n)e(n), \tag{4.3}$$

where we define the gain factor $g(n)$ as the root mean square (RMS) value of $r(n)$ within a window of length L_{g} samples, and

$$g(n) = \left[\frac{1}{L_{\mathrm{g}}} \sum_{m=0}^{L_{\mathrm{g}}-1} r^2 \left(n - m + \tfrac{L_{\mathrm{g}}-1}{2} \right) \right]^{\frac{1}{2}}.$$

We use a short window duration of 2 ms (corresponding to L_g samples), which maintains the perceptually important short-term temporal waveform envelope.

To determine the segmentation of the speech signal into voiced and non-voiced components, we use the PRAAT implementation of an autocorrelation-based pitch tracking algorithm, which detects the pitch pulses in the speech signal with a local cross-correlation value maximization [77].

4.2.2. Watermark Signal Spectral Shaping

The pulse shaping requirements described in Section 4.1.5.2 and the constraints on N, M, f_{L}, f_{H} and y_p given a certain transmission channel passband and a certain sampling frequency f_s are theoretically demanding and difficult to meet in practice. Even though the topic is well explored in literature (e.g., [81, 82, 83]), there is at present no automatic procedure to obtain 'good' parameter sets which fulfill all stated requirements such as being bandwidth-efficient and free of inter-symbol interference. Thus, we now provide suitable parameter sets and interpolation filters for three application scenarios of practical relevance, namely a lowpass channel, a narrowband telephony channel, and an aeronautical radio channel scenario. We do so for a sampling frequency $f_s = 8000\,\mathrm{Hz}$.

The derived reconstruction pulse shapes $y_{b,p}(t)$ are continuous-time and of infinite length. To implement the pulse shapes as digital filters, $y_{b,p}(t)$ is sampled at the time instants $t = nT_s$, $n \in \mathbb{Z}$, and truncated or windowed.

4.2.2.1. Lowpass Channel

For a lowpass channel with bandwidth W_C and choosing M, N and f_s according to $W_C = \frac{M}{N}\frac{f_s}{2}$, one can obtain the desired signal $w(t)$ that meets all requirements with the interpolation function

$$y_{b,p}(t) = \frac{(-1)^{bM}}{\frac{\pi f_s}{N}(t - t_p - bNT_s)} \cdot \frac{\prod_{q=1}^{M} \sin\left(\frac{\pi f_s}{N}(t - t_q)\right)}{\prod_{q=1,q\neq p}^{M} \sin\left(\frac{\pi f_s}{N}(t_p - t_q)\right)},$$

which is a general formula for bandwidth-limited lowpass signals and achieves the Nyquist rate [85]. In contrast, using for example a simple sinc function for $y_{b,p}(t)$ would result in either ISI or the sample rate not being equal to f_s.

4.2.2.2. Narrowband Telephony Channel

For a narrowband telephony channel with a passband from 300 Hz to 3400 Hz we select $M = 4$ and $N = 6$ and define the reconstruction pulse shape

$$\begin{aligned} y_{b,p}(t) = y_p(t - bNT_s) = y_p(t^{(b)}) = \\ \operatorname{sinc}\left(2\pi\frac{f_s}{12}\left[t^{(b)} - (p-1)T_s\right]\right) \\ \cdot \sin\left(2\pi\frac{f_s}{12}\left[t^{(b)} - \left(p - 1 + 2\left\lceil\tfrac{p}{2}\right\rceil\right)T_s\right]\right) \\ \cdot \cos\left(2\pi\frac{f_s}{4}\left[t^{(b)} - (p-3)T_s\right]\right), \\ t^{(b)} := t - bNT_s, \qquad \operatorname{sinc}(x) = \frac{\sin(x)}{x}, \end{aligned} \tag{4.4}$$

to obtain a watermark signal w with a bandwidth from 666 Hz to 3333 Hz. Figure 4.4 shows this by illustration. This passband width $W_C = 2666$ Hz is as small as the Nyquist bandwidth that is required for the chosen symbol rate $\frac{M}{N}f_s$, which is only achievable for distinct parameter combinations. Figure 4.5 shows the M pulse shapes and demonstrates how each pulse creates ISI only in the $N - M$ non-information carrying samples. This is achieved by carefully selecting the frequencies and phases of the terms in (4.4) such that at each information-carrying sampling position a zero-crossing of one of the terms occurs.

Ayanoglu *et al.* [84] define similar parameter and signal sets for channels band-limited to 0 Hz–3500 Hz ($M = 7$, $N = 8$) and 500 Hz–3500 Hz ($M = 6$, $N = 8$).

4.2.2.3. Aeronautical Radio Channel

Given the particular passband width and position of the aeronautical radio channel (300 Hz to 2500 Hz), using the aforementioned approach did not lead to a configuration that would achieve the Nyquist rate.

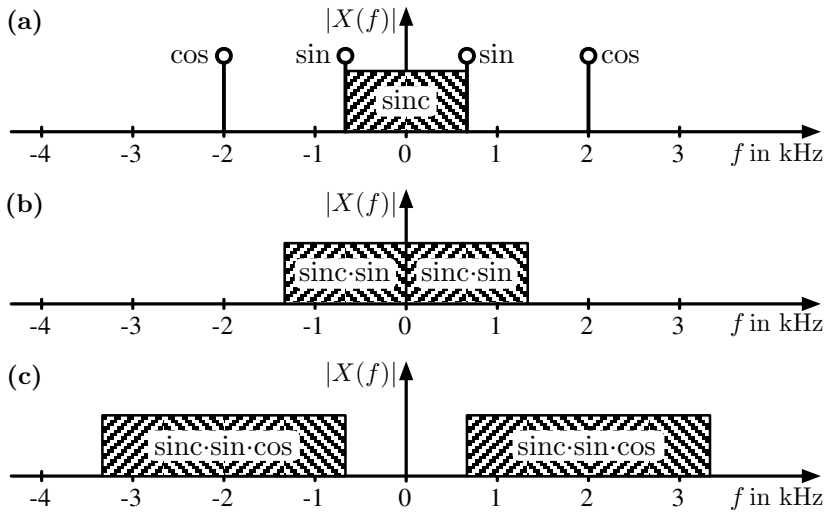


Figure 4.4.: Generation of a passband watermark signal for a telephony channel using (4.4). Frequency domain representation **(a)** of the individual sinc, sin, and cos terms, **(b)** of the product of the first two terms, and **(c)** of the product of all three terms in (4.4), resulting in the desired bandwidth and position.

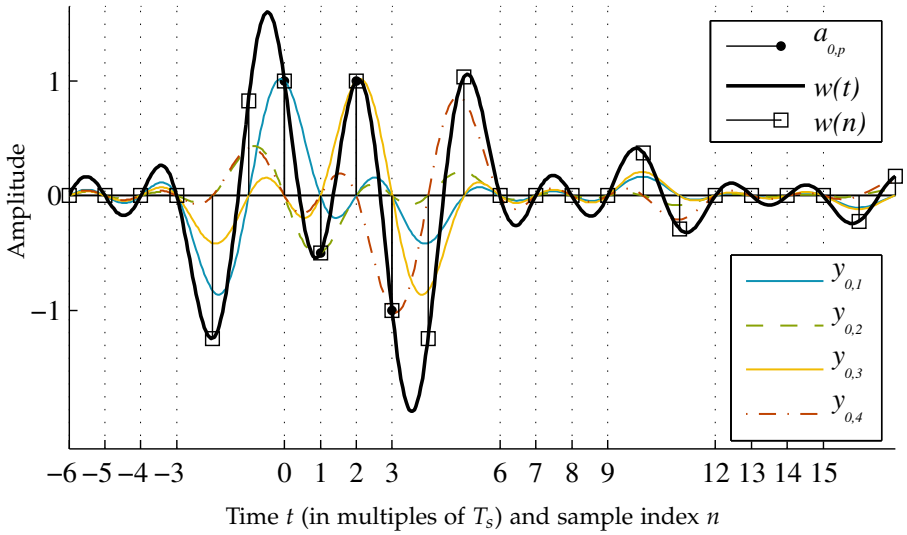


Figure 4.5.: Pulse shapes for zero ISI between the data-carrying samples $a_{b,p}$. The example data $a_{0,p=1\ldots4} = 1, -\frac{1}{2}, 1, -1$ is located at $t_p = (p-1)T_s$. By design, the ISI induced by the band-limitation falls into the spectral shaping samples at $t = b \cdot 6T_s + 4T_s$ and $t = b \cdot 6T_s + 5T_s$, with $b \in \mathbb{Z}$.

However, an efficient reconstruction pulse set with a watermark bandwidth from 500 Hz to 2500 Hz can be created using Kohlenberg's second-order sampling [88]. We omit the exact description of the method and only show its practical application: Using the notation of [88], we define $W_0 = 500\,\text{Hz}$ and $W = 2000\,\text{Hz}$, which results in the parameter $r = 1$ and fulfills $\frac{2W_0}{W} < r < \frac{2W_0}{W} + 1$ [81]. We select $a = a_1 = a_2 = \frac{1}{W}$, which represents every fourth sample in terms of the sample rate $f_s = 8000\,\text{Hz} = 4W$. The phase shift for the second-order sampling is chosen to be $k = \frac{1}{4W}$, which corresponds to a shift of one sample in terms of f_s. This results in $M = 2$ consecutive information carrying samples (indexed by $p = 1, 2$) in each block of $N = 4$ samples, and the symbol rate achieves the Nyquist rate. The desired passband signal w is then given similar to (4.1) and (4.4) by

$$w(t) = \sum_{b=-\infty}^{\infty} \left[a_{b,1} s(t^{(b)}) + a_{b,2} s(k - t^{(b)}) \right] \tag{4.5}$$

with $t^{(b)} = t - bNT_s$ and

$$s(t) = \frac{\cos\left[2\pi(W_0 + W)t - (r+1)\pi Wk\right] - \cos\left[2\pi(rW - W_0)t - (r+1)\pi Wk\right]}{2\pi Wt \sin\left[(r+1)\pi Wk\right]} + \frac{\cos\left[2\pi(rW - W_0)t - r\pi Wk\right] - \cos\left[2\pi W_0 t - r\pi Wk\right]}{2\pi Wt \sin\left[r\pi Wk\right]}$$

as defined in [88, Eq. 31, with $k \equiv K$].

4.3. Experiments

This section presents experimental results demonstrating the feasibility, capacity, and robustness of the proposed method.

4.3.1. Experimental Settings

Motivated by legacy telephony and aeronautical voice radio applications, the system was evaluated using a narrowband speech configuration and various simulated transmission channels.

4.3.1.1. Watermark Embedding

As input signal we used ten randomly selected speech utterances from the ATCOSIM corpus [5]. They were chosen by a script such that there was one utterance per speaker (six male and four female) and such that each utterance had a length between 5 and 7 seconds, resulting in a total of 57 s of speech. The signal was resampled to $f_s = 8000\,\text{Hz}$, analyzed with an LP order $P = 10$, and the watermark embedded in a frequency band from 666 Hz to 3333 Hz using $M = 4$ and $N = 6$ (Section 4.2.2.2). In each active frame of length $L_\text{D} = 180$ samples, $L_\text{WM} = 34$ symbols were allocated for watermark data, $L_\text{Tr} = 72$ for training symbols, $L_\text{Ac} = 7$ for active frames markers, $L_\text{ID} = 7$ for frame

indices, and $L_{\text{Sh}} = 60$ symbols for spectral shaping. For all but the spectral shaping symbols we used unit length binary symbols with alphabet $\{-1; 1\}$.

There is a large number of possibilities for the above parameter choices, which form a trade-off between data rate and robustness. The parameters were picked manually such that the corresponding subsystems (equalization, frame synchronization, detection, ...) showed satisfactory performance given the channel model in Table 4.1. The experimental settings are as such not carefully optimized and serve as illustrative example, only.

4.3.1.2. Transmission Channels

The watermarked speech signal $\hat{s}_{\text{FB}}(n) = \hat{s}_{\text{PB}}(n) + s_{\text{SB}}(n)$ was subjected to various channel attacks. Motivated by the application to telephony and aeronautical voice radio communication, we simulated the following transmission channels:

1. Ideal channel (no signal alteration).
2. Filtering with a linear phase FIR digital bandpass filter of order $N = 200$ and a passband from 300 Hz to 3400 Hz.
3. Sinusoidal amplitude modulation (flat fading) with a modulation frequency of $f_{\text{AM}} = 3$ Hz and a modulation index (depth) of $h_{\text{AM}} = 0.5$.
4. Additive white Gaussian noise (AWGN) with a constant segmental SNR of 30 dB, using a window length of 20 ms.
5. Filtering with an FIR linear phase Intermediate Reference System (IRS) transmission weighting filter of order $N = 150$ as specified in ITU-T P.48 [89] and implemented in ITU-T STL G.191 [90].
6. Filtering with a measured aeronautical voice radio channel response from the TUG-EEC-Channels database, an FIR filter with non-linear and non-minimum phase [4].
7. Filtering with an IIR allpass filter with non-linear and non-minimum phase and z-transform
$$H(z) = \frac{1 - 2z^{-1}}{1 - 0.5z^{-1}}.$$
8. A *combination* of AWGN, sinusoidal amplitude modulation and aeronautical voice radio channel filtering, each as described above.

The magnitude responses of the bandpass, IRS and aeronautical channel filters are shown in Figure 4.6 with respect to the watermark embedding band. While for the aeronautical channel the applied embedding bandwidth is too large and should instead be chosen according to the specified channel bandwidth as in Section 4.2.2.3, the shown configuration is a worst-case scenario and tests if the watermark detector can handle the high attenuation within the embedding band. The simulated transmission channels are

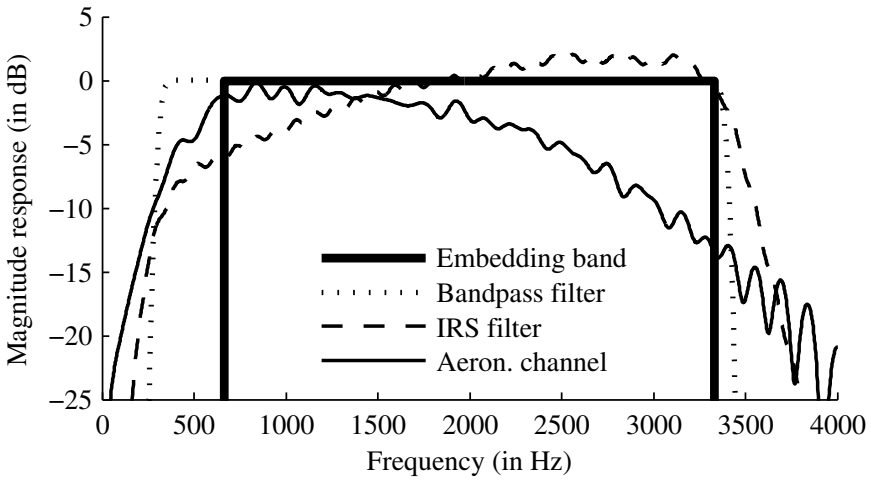


Figure 4.6.: Magnitude responses of the bandpass, IRS and aeronautical transmission channel filters, in comparison with the watermark embedding band.

in the digital domain. Robustness against desynchronization and D/A-A/D conversion is evaluated independently in Chapter 5.

We set the SNR of the AWGN channel to a level that results in an acceptable overall BER, and the SNR is as such related to the embedding parameters of Section 4.3.1.1. It is likely that a real-world aeronautical channel has at times a worse SNR, and different parameter settings, for example a higher data symbol dimensionality, might be preferable in the aeronautical application.

4.3.1.3. Watermark Detection

We used an adaptive RLS equalizer filter with $K = 11$ taps, $\lambda = 0.7$ and $\alpha = 0.5$. The number of filter taps is a trade-off between the channel's memory and the adaptive filter's tracking ability in light of the fast time-variation of the vocal tract filter and the radio channel. In the noise-less and time-invariant case, an (RLS) FIR filter with $K = 11$ taps can constitute the perfect inverse of a LP all-pole filter of order $P = 10$. In practice, the LP filter is time-variant and the observation distorted and noise-corrupted by the radio transmission channel, which results in imperfect inversion. In contrast to the description in Section 4.1.7.2, in the experiments presented herein we updated the inverse correlation matrix $\mathbf{P}(n)$ only when a training symbol was available.

4.3.2. Overall System Evaluation

4.3.2.1. Robustness

We evaluated the overall watermarking system robustness (including frame synchronization) in the presence of the channel attacks listed in the previous subsection and using the input signal defined therein. We measured the robustness in terms of raw bit

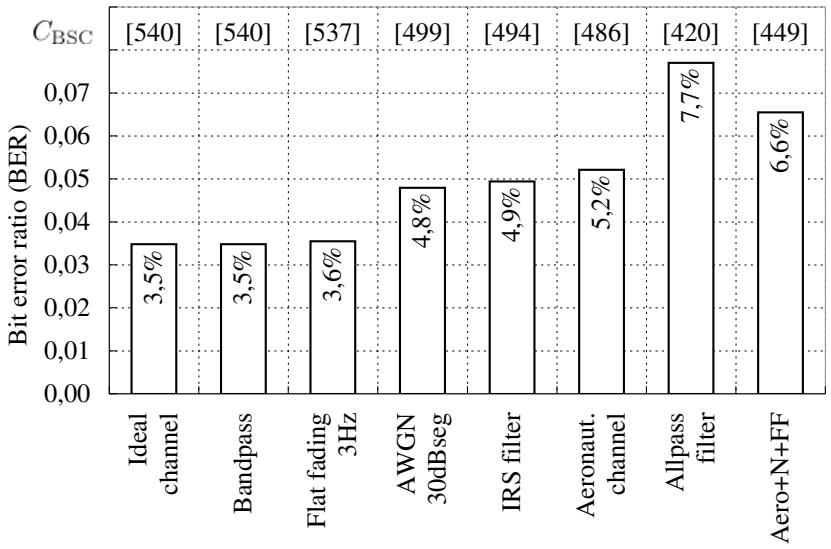


Figure 4.7.: Overall system robustness in the presence of various transmission channel attacks at an average uncoded bit rate of 690 bit/s. The numbers in the top row denote the corresponding BSC capacity C_{BSC} in bit/s.

error ratio (BER) in the detected watermark data, without any forward error correction coding being applied. The results are shown in Figure 4.7.

4.3.2.2. Data Rate

Out of the 57 s of speech a total of 31 s or $\gamma = 53\,\%$ were classified as non-voiced, resulting with the bit allocation of Section 4.3.1.1 in a measured average payload data rate of $R \approx 690\,\mathrm{bit/s}$ in terms of uncoded binary symbols m_{WM}. The rate R is approximately $\frac{f_s \gamma L_{\mathrm{WM}}}{L_{\mathrm{D}}}$ times a factor that compensates for the unused partial frames in non-voiced segments as shown in Figure 4.2.

Given the transmission rate as well as the BERs shown in Figure 4.7, it is possible to express the achievable watermark capacity at which error-free transmission is possible by considering the payload data channel as a memoryless binary symmetric channel (BSC). The channel capacity C_{BSC} as a function of the given rate R and BER p_{e} is (e.g. [91])

$$C_{\mathrm{BSC}}(R, p_{\mathrm{e}}) = R\left(1 + p_{\mathrm{e}} \log_2\left(p_{\mathrm{e}}\right) + \left(1 - p_{\mathrm{e}}\right) \log_2\left(1 - p_{\mathrm{e}}\right)\right) \tag{4.6}$$

in bit/s and asymptotically achievable with appropriate channel coding.

4.3.2.3. Listening Quality

The listening quality of the watermarked speech signal $\hat{s}_{\mathrm{FB}}(n)$ was evaluated objectively using ITU-T P.862 (PESQ) [92], resulting in a mean opinion score MOS-LQO of 4.05 (on a scale from 1 to 4.5). Audio files with the original and the watermarked speech signals

$s_{\mathrm{FB}}(n)$ and $\hat{s}_{\mathrm{FB}}(n)$ are available online [93]. Informal listening tests showed that the watermark is hardly perceptible in the presence of an ideal transmission channel, and not perceptible in the presence of the simulated radio channel with AWGN, amplitude modulation and channel filtering as defined above.

4.3.3. Evaluation of Alternative Detection Schemes

To gain insight into the proposed RLS channel equalization, we compared its performance to two alternative adaptive filtering strategies by evaluating their robustness in line with Section 4.3.2.1.

4.3.3.1. Detection Using LP Analysis

We previously proposed to determine the adaptive filter coefficients in the detector using LP analysis, like in the embedding [3]. To achieve sufficiently similar coefficients in the embedder and the detector, in the presence of a bandpass channel the LP analysis in the embedder must then be based on $s_{\mathrm{PB}}(n)$ instead of $s_{\mathrm{FB}}(n)$.[1] In comparison to Figure 4.7 the BER approximately triples for the ideal, bandpass, noise and flat fading channels, and for the IRS and the non-linear phase channels the BER is 50 % and watermark detection fails altogether.

4.3.3.2. Detection Using RLS Equalization

Using the proposed RLS equalization scheme, the watermark data can be detected for all channels. The BERs are shown in Figure 4.7.

4.3.3.3. Detection Using Embedder's Predictor Coefficients

Using the *embedder*'s predictor coefficients $\mathbf{c}(n)$ for adaptive filtering in the detector is a purely hypothetical experiment, since in most applications the coefficients of the embedder are not available in the detector. However, it serves as an upper bound on how well the RLS adaptive filter would theoretically do when neglecting the transmission channel and perfectly inverting the time-variant all-pole filtering $H_{\mathbf{c}}$. The obtained BERs are lower by approximately one magnitude compared to the RLS results in Figure 4.7, but only so if the transmission channel is linear phase. In the case of the non-linear phase channels the BER is 50 %, and watermark detection fails.

4.4. Discussion

This section presents a qualitative summary and interpretation of the experimental findings.

[1] Given that $s'_{\mathrm{PB}}(n)$ is a passband signal, it might seem beneficial to use *selective* linear prediction (SLP) [94] to estimate the coefficients. SLP is, however, only a spectral estimation technique that does not yield the actual filter coefficients required in this application.

4.4.1. Comparison with Other Schemes

First and foremost, no other schemes are known at present that are complete (in the sense of most practical issues such as synchronization being addressed), are robust with respect to the transmission channel model of Section 4.1.2, and have an embedding rate that is comparable to the one shown in Section 4.3.2.2. It is the combination of these three aspects that differentiates our proposal from current state-of-the-art methods and our own previous work.

For a numerical comparison of our results with the reported performance of other methods we use as measure C_{BSC}/W, based on the number of embedded watermark bits R (in terms of C_{BSC} using the reported BER and (4.6)) and per kHz of embedding or channel bandwidth W. Table 4.2 indicates that our method outperforms most of the current state-of-the-art speech watermarking methods. It is, however, difficult to obtain an objective ranking, because almost all methods are geared towards different channel attacks and applications, and in general a large number of factors contribute to the 'goodness' of each system. While the method presented in [32] appears to perform similar to our method, it assumes a time-invariant transmission channel and requires before every transmission a dedicated non-hidden 4 s long equalizer training signal. This makes the method impractical for the considered aeronautical application.

Table 4.2 also includes our previous work and shows that there are multiple options for adapting our approach to different channel conditions. We have previously applied various measures, such as the use of a different data symbol dimensionality, the addition of a watermark floor (that is, a fixed minimum watermark signal gain g), or a pre-processing of the host signal, to adjust the trade-off between perceptual fidelity, capacity and robustness [2].

4.4.2. Comparison with Channel Capacity

In the following we compare the achieved embedding rate with the theoretical capacity derived in Chapter 3. In our practical scheme, we embed 690 bit/s at 4.79 % BER, which corresponds to a BSC capacity of 500 bit/s. With a maximum symbol transmission rate or Nyquist rate $2W_C = 5333$ symbols/s and $C_{\text{W, Phase}} \approx 3$ bit/symbol (using (2.2) with an SNR of 30 dB as applied in Section 4.3.2.1), the theoretical capacity evaluates to $C_{\text{W}} = 16000$ bit/s, or $C_{\text{W}} = 8500$ bit/s considering embedding in non-voiced speech, only.

In non-voiced speech (comprising unvoiced speech and pauses), we achieve the Nyquist rate $2W_C$ with all methods presented in Section 4.2.2. However, we embed only one bit per symbol instead of three, because our current method does not account for the time-variation and the spectral non-uniformity of the SNR that result from the spectral characteristics of the host signal. Table 4.3 summarizes these and other factors that contribute to the capacity gap. The given rate losses are multiplicative factors derived from the experimental settings of Section 4.3.1.1 and the observed system performance. The table shows where there is room for improvement, but also shows that (2.2) is an over-estimation since it does not account for the perceptually required temporally constrained embedding, the time-variant and colored spectrum of the host

Table 4.2.: Comparison With Reported Performance of Other Schemes

Method	Ref.	R (bit/s)	BER	C_{BSC} (bit/s)	W (kHz)	C_{BSC}/W (bit/(s kHz))	SNR	Simultan. ch. attacks: LP or BP	Nonlin. phs.	Fading ch.	DA-AD cnv.
Proposed Method											
Phase Watermarking	4.3.	690	4,8%	499	2,7	187	30 dB_{sg}	X		X	
Phase Watermarking	4.3.	690	6,6%	449	2,7	168	30 dB_{sg}	X	X	X	
Ph. WM (vector sym.)	[2]	300	9,0%	169	4,0	42	10 dB				X
Ph. WM (vector sym.)	[2]	300	2,0%	258	4,0	64	15 dB				X
Ph. WM (scalar sym.)	[2]	2000	11,0%	1000	4,0	250	20 dB				X
Ph. WM (scalar sym.)	[2]	2000	2,5%	1663	4,0	416	30 dB				X
Ph. WM (vector sym.)	[3]	130	0,0%	130	4,0	32	∞				
Ph. WM (scalar sym.)	[3]	2000	0,8%	1866	4,0	466	∞				
Alternative Speech Methods											
Spread spectrum	[28]	24	0,0%	24	2,8	9	20 dB	X		X	
Spread spectrum	[27]	800	28,4%	111	4,0	28	∞	X			X
QIM of AR coeff.	[29]	4	3,0%	3	6,0	1	∞				
QIM of AR coeff.	[37]	24	n/a	24	8,0	3	n/a	X			
QIM of AR residual	[33]	300	n/a	300	3,1	97	n/a				
QIM of pitch	[31]	3	1,5%	3	4,0	1	n/a				
QIM of DHT coeff.	[32]	600	0,0%	600	3,0	200	35 dB				
QIM of DHT coeff.	[32]	600	0,1%	600	3,0	200	$V.56_{bis}$	X	X		
Mod. of partial traj.	[35]	200	1,0%	184	10,0	18	∞				
Repl. of maskees	[39]	348	0,1%	344	4,0	86	25 dB				
Alternative Audio Methods											
QIM of DCT of DWT	[26]	420	0,0%	418	2,0	209	∞	X			
QIM of DCT of DWT	[26]	420	0,0%	418	22,5	19	10 dB				
Allpass phase mod.	[45]	243	10,0%	129	3,0	43	∞	X			
Allpass phase mod.	[45]	243	2,0%	209	7,0	30	5 dB				

Table 4.3.: Causes for Capacity Gap and Attributed Rate Losses

Cause	Loss in Rate
No embedding in voiced speech	47 %
Embedding of training symbols	60 %
Embedding of frame headers	29 %
Use of fixed frame position grid	14 %
Embedding in non-white host signal	67 %
Imperfect tracking of time-variation	28 %

signal, and the filtering characteristic of the transmission channel.

4.4.3. Equalization and Detection Performance

Three different methods to obtain the filter coefficients for the adaptive filtering in the detector were evaluated. All methods are invariant against the bandpass filtering and the flat fading transmission channel. Additive WGN at a moderate SNR does have an influence on the bit error ratio but does not disrupt the over-all functioning of the system.

The hypothetical experiment of using the embedder coefficients in the detector shows that approximating the embedder coefficients is a desirable goal only if the channel is linear phase. Using linear prediction in the detector results in a bit error ratio of approximately 10 % even in the case of an ideal channel. Compared to our previous results [2] this is a degradation, which is caused by the band-limited embedding introduced herein.

The proposed RLS adaptive filtering based detection is the only method robust against all tested transmission channels. It can compensate for a non-flat passband and for non-linear and non-minimum phase filtering due to the embedded training symbols. Over a channel that combines AWGN, flat fading, and a measured aeronautical radio channel response, the method transmits the watermark with a data rate of approximately 690 bit/s and a BER of 6.5 %, which corresponds with (4.6) to a BSC capacity of 450 bit/s. The same RLS method is applied in [32] for a time-invariant channel using a 4 s long initial training phase with a clearly audible training signal. In contrast, we embed the training symbols as inaudible watermarks and perform on-the-fly equalizer training and continuous tracking of the time-variant channel.

A further optimization seems possible by developing an equalization scheme that is not solely based on the embedded training symbols, but that also incorporates available knowledge about the source signal and the channel model, such as the spectral characteristics.

4.5. Conclusions

The experimental results show that it is possible to embed a watermark in the phase of non-voiced speech. Compared to quantization-based speech watermarking, the principal advantage of our approach is that the theoretical capacity does not depend on an embedding distortion to channel noise ratio, but depends only on the host signal to channel noise ratio. We presented a proof-of-concept implementation that considers many practical transmission channel attacks. While not claiming any optimality, it embeds 690 bit/s with a BER of 6.6 % in a speech signal that is transmitted over a time-variant non-linear phase bandpass channel with a segmental SNR of 30 dB and a bandwidth of approximately 2.7 kHz.

The large gap between the presented implementation and the theoretical capacity shows that there is a number of areas where the performance can be further increased. For example, a decision feedback scheme for the RLS equalization, or a *joint* synchronization, equalization, detection and decoding scheme could increase the reliability of the watermark detector. On a larger scale, one could account for the non-white and time-variant speech signal spectrum with an adaptive embedding scheme, and apply the same embedding principle on all time-segments of speech for example using a continuous noise-like versus tonal speech component decomposition [95].

Chapter 5

Watermark Synchronization

This chapter discusses different aspects of the synchronization between watermark embedder and detector. We examine the issues of timing recovery and bit synchronization, the synchronization between the synthesis and the analysis systems, as well as the data frame synchronization. Bit synchronization and synthesis/analysis synchronization are not an issue when using the adaptive equalization-based watermark detector of Chapter 4. For the simpler linear prediction-based detector we present a timing recovery mechanism based on the spectral line method which achieves near-optimal performance.

Using a fixed frame grid and the embedding of preambles, the information-carrying frames are detected in the presence of preamble bit errors with a ratio of up to 10 %. Evaluated with the full watermarking system, the active frame detection performs near-optimal with the overall bit error ratio increasing by less than 0.5 %-points compared to ideal synchronization.

Parts of this chapter have been published in K. Hofbauer and H. Hering, "Noise robust speech watermarking with bit synchronisation for the aeronautical radio," in *Information Hiding*, ser. Lecture Notes in Computer Science. Springer, 2007, vol. 4567/2008, pp. 252–266.
K. Hofbauer, G. Kubin, and W. B. Kleijn, "Speech watermarking for analog flat-fading bandpass channels," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, revised and resubmitted.

"Synchronization is the process of aligning the time scales between two or more periodic processes that are occurring at spatially separated points"[96]. It is a multilayered problem, and certainly so for an analog radio channel. We discuss in this chapter the different aspects of synchronization between the watermark embedder and the watermark detector, and present a practical synchronization scheme. Section 5.1 presents related work and our approach to the different levels of synchronization from the shortest to the longest units of time. In Section 5.2 we describe an implementation of the synchronization scheme, which is tailored to the watermarking algorithm of Chapter 4. Experimental results are shown and discussed in Section 5.3.

5.1. Theory

Due to the analog nature of the transmission channel and the signal-dependent embedding of the data in the speech signal, a watermark detector as shown in Figure 4.3 needs to estimate 1) the correct sampling time instants, 2) the transformations applied to the data, and 3) the position of the data-carrying symbols.

5.1.1. Timing Recovery and Bit Synchronization

When transmitting digital data over an analog channel, the issue of bit or symbol synchronization in-between digital systems arises. The digital sampling clocks in the embedder and detector are in general slowly time-variant, have a slightly different frequency, and have a different timing phase (which is the choice of sampling instant within the symbol interval). Therefore, it is necessary that the detector clock synchronizes itself to the incoming data sequence or compensates for the sampling phase shift. Although bit synchronization is a well explored topic in communications, it is still a major challenge in most modern digital communication systems, and even more so in watermarking due to very different conditions. Bit synchronization can be achieved with data-aided or non-data-aided methods, and we will use one or the other depending on the used watermark detector structure.

5.1.1.1. Data-Aided Bit Synchronization

In data-aided bit synchronization, the transmitter allocates a part of the available channel capacity for synchronization. It is then possible to either transmit a clock signal, or to interleave the information data with a known synchronization sequence [65]. This enables simple synchronizer designs, but also reduces the channel capacity available for information transmission.

Given the RLS equalization based watermark detection presented in Section 4.1.7.2 (which is necessary for a channel with linear transmission distortion), there is already a known sequence used for equalizer training embedded in the signal. While this training sequence could also be used to drive a data-aided bit synchronization method, this is in fact not necessary for the RLS based watermark detector. The RLS equalizer inherently compensates for a timing or sampling phase error, as long as the error is smaller than the RLS filter length.

In the aeronautical application discussed in Chapter 6, the radio-internal frequency references are specified to be accurate within ± 5 ppm (parts per million) for up-to-date airborne transceivers and ± 1 ppm for up-to-date ground transceivers [97]. Assuming a sampling frequency of 8 kHz and a worst-case frequency offset of 6 ppm, this leads to a timing phase error of 0.048 samples per second, which accumulates over time but is less than one sample (and well below the RLS filter length) given the short utterance durations in air traffic control.

5.1.1.2. Non-Data-Aided Bit Synchronization

We discussed in Section 4.3.3 the use of linear prediction (LP) instead of the RLS equalizer as adaptive filter in the watermark detector, in case the transmission channel has no linear filtering distortion. In this case, no equalizer training sequence is embedded, and payload data can be transmitted, instead. Since the LP error filter does not compensate for timing phase errors, a sampling timing recovery method is required, which is ideally non-data-aided in order to obtain a high payload data rate.

In non-data-aided bit synchronization the detector achieves self-synchronization by extracting a clock-signal from the received signal. A wide range of methods exist, including early-late gate synchronizers, minimum mean-square-error methods, maximum likelihood methods and spectral line methods [98, 65]. For our watermarking system, we use the highly popular nonlinear spectral line method because of its simple structure and low complexity [65, 86]. It belongs to the family of deductive methods. Based on higher-order statistics of the received signal, it derives from the received signal a timing tone whose frequency equals the symbol rate and whose zero crossings indicate the desired sampling instants.

For non-voiced speech our watermarking scheme is essentially a pulse amplitude modulation (PAM) system with a binary alphabet. To demonstrate the basic idea of the spectral line method, we assume that the received analogue watermark signal is a baseband PAM signal $R(t)$ based on white data symbols with variance σ_a^2. Then $R(t)$ has zero mean but non-zero higher moments that are periodic with the symbol rate. With $p(t)$ being the band-limited interpolation pulse shape, the second moment of $R(t)$ is [86]

$$E\left[|R(t)|^2\right] = \sigma_a^2 \sum_{m=-\infty}^{\infty} |p(t-mT)|^2 ,$$

which is periodic with the symbol period T. A narrow bandpass filter at $f_s = \frac{1}{T}$ is used to extract the fundamental of the squared PAM signal $R^2(t)$, which results in a timing tone with a frequency corresponding to the symbol period and a phase such that the negative zero crossings are shifted by $\frac{T}{4}$ in time compared to the desired sampling instants [65]. We compensate this phase shift with a Hilbert filter.

Due to the noisy transmission channel, and even though most of the noise is filtered out by the narrow bandpass filter at $f_s = \frac{1}{T}$, the timing tone is noise-corrupted. This leads to a timing jitter, which we reduce using a phase-locked loop that provides a stable single-frequency output without following the timing jitter.

5.1.2. Synthesis and Analysis System Synchronization

In principle, the signal analysis in the watermark detector needs to synchronize to the signal synthesis in the watermark embedder. In order to perfectly detect the watermark signal $w(n)$ (see Figure 4.1), the LP analysis, the voiced/unvoiced segmentation, as well as the gain extraction in the detector would have to be perfectly in sync with the same processes in the watermark embedder. This is difficult to achieve because these processes are signal-dependent and the underlying signal is modified by a) the watermarking process and b) the transmission channel. In the context of our watermarking system it is only a matter of definition if one considers these issues as a synchronization problem, or if they are considered as part of the hidden data channel as shown in 4.1. We address these issues with different approaches.

LP Frame Synchronization

When using an LP-based watermark detector, one might intuitively assume that the block boundaries for the linear prediction analysis in the embedder and detector have to be identical. However, using LP parameters as given in Chapter 4 and real speech signals, the predictor coefficients do not change rapidly in between the update interval of 2 ms. As a consequence, a synchronization offset in the LP block boundaries is not an issue. In [2] we also showed experimentally that the bit error rate is not affected when the LP block boundaries in the detector are offset by an integer number of samples compared to the embedder.

When using the RLS-based watermark detector, LP frame synchronization does not exist since the RLS equalizer operates on a per sample basis.

Adaptive Filtering and Gain Modulation Mismatch

The mismatch between the LP synthesis filter in the embedder and the adaptive filtering in the detector, as well as the the gain estimation mismatch between watermark embedder and detector are mostly a result of the transmission channel attacks. We address this mismatch in two ways. First, we reduce the mismatch using the embedded training sequences and the RLS equalizer in the receiver. Second, we accept that there is a residual mismatch such that we are unable to exactly recuperate the watermark signal $w(n)$. Therefore, we only use a binary alphabet for $w(n)$ to achieve sufficient robustness.

Voiced/Unvoiced Segmentation

Due to the transmission channel attacks outlined in 4.1 it appears difficult to obtain an identical sample-accurate voiced/unvoiced segmentation in the embedder and the detector. While the embedder knows about the availability of the hidden watermark channel (i.e., the host signal being voiced or non-voiced) and embeds accordingly, the watermark channel appears to the detector as a channel with random insertions, deletions and substitutions. We deal with this issue in combination with the frame synchronization described in the next subsection.

5.1.3. Data Frame Synchronization

In this subsection we address the localization of the information-carrying symbols in the watermark detector. Due to the host signal dependent watermark channel availability, the hidden data channel appears to the detector as a channel with insertions, deletions and substitutions (IDS). Various schemes have been proposed for watermark synchronization, including embedding in an invariant domain, embedding of preambles or pilot sequences, host signal feature based methods, the use of IDS-capable channel codes, the use of convolution codes with multiple interconnected Viterbi decoders, and exhaustive search (cf. [99, 100, 101, 102]).

We use a fixed frame grid and an embedding of preambles, since this allows the exploitation of the watermark embedder's knowledge of the watermark channel availability, and does not require a synchronized voicing decision in the detector. As outlined in Section 4.1.5.1, a preamble is embedded in frames that carry a watermark, so-called active frames, and consists of 1) a fixed marker sequence m_{Ac} that identifies active frames and 2) a consecutive frame counter m_{ID} used for addressing or indexing among the active frames.

To achieve synchronization at the watermark detector, a threshold-based correlation detection of the active frame markers m_{Ac} over a span of multiple frames detects the frame grid position and as such the overall signal delay. After the adaptive equalization of the received signal as described in Section 4.1.7, there remain errors in the detection of the preambles due to channel attacks. Therefore, the active frame detection is based on a dynamic programming approach that detects IDS errors and minimizes an error distance that incorporates the active frame markers m_{Ac}, the frame indices m_{ID}, and the training symbols m_{Tr}.

Frame synchronization is considered as a secondary topic in the scope of this thesis and does not significantly influence the overall performance. Alternative schemes, which perform similarly, are possible. An introduction to frame synchronization is provided in [103].

5.2. Implementation

5.2.1. Bit Synchronization

For the RLS-based watermark detector data-aided bit synchronization is inherently achieved by the RLS equalizer. Thus, we focus in the following on the non-data-aided bit synchronization for the LP-based watermark detector using the spectral line method.

Spectral Line Synchronization Figure 5.1 shows the block diagram of the proposed spectral line synchronization system.[1] The received watermarked analog signal $s(t)$ is oversampled by a factor k compared to the original sampling frequency f_s. In the

[1] Whether the proposed structure could be implemented even in analog circuitry could be subject of further study.

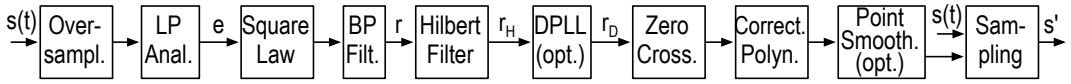


Figure 5.1.: Synchronization system based on spectral line method.

later simulations a factor $k = 32$ is used. Values down to $k = 8$ are possible at the cost of accuracy. The linear prediction residual $e(n)$ of the speech signal $s(t)$ is computed using LP analysis of the same order $P = 10$ as in the embedder and with intervals of equal length in time. This results in a window length of $k \cdot 160$ samples and an update interval of $k \cdot 15$ samples after interpolation.[2]

We exploit the fact that the oversampled residual shows some periodicity with the embedding period $T = \frac{1}{f_s}$ due to the data embedding at these instances. We extract the periodic component $r(n)$ at f_s from the squared residual $(e(n))^2$ with an FIR bandpass filter with a bandwidth of b=480 Hz centered at f_s. The output $r(n)$ of the bandpass filter is a sinusoid with period T, and is phase-shifted by $\frac{\pi}{2}$ with an FIR Hilbert filter resulting in the signal $r_H(n)$. The Hilbert filter can be designed with a large transition region given that $r(n)$ is a bandpass signal.

The zero-crossings of $r_H(n)$ are determined using linear interpolation between the sample values adjacent to the zero-crossings. The positions of the zero-crossings on the rising slope are a first estimate of the positions of the ideal sampling points of the analog signal $s(t)$. It was found that the LP framework used in the simulations introduces a small but systematic fractional delay which depends on the oversampling factor k and results in a timing offset. We found that this timing offset can be corrected using a third-order polynomial $t_\triangle = a_0 + a_1 k^{-1} + a_2 k^{-2} + a_3 k^{-3}$. The coefficients a_i have been experimentally determined to be $a_0 = 0$, $a_1 = 1.5$, $a_2 = -7$ and $a_3 = 16$.

Since the estimated sampling points contain gaps and spurious points, all points whose distance to a neighbor is smaller than 0.75 UI (unit interval, fraction of a sampling interval $T = \frac{1}{f_s}$) are removed in a first filtering step. In a second step all gaps larger than 1.5 UI are filled with new estimation points which are based on the position of previous points and the observed mean distance between the previous points. The received analog signal $s(t)$ is again sampled, but instead of with a fixed sampling grid it is now sampled at the estimated positions. The output is a discrete-time signal with rate f_s, which is synchronized to the watermark embedder and which serves as input to the watermark detector.

Digital Phase-Locked Loop The bit synchronization can be further improved by the use of a digital phase-locked loop (DPLL). The DPLL still provides a stable output in the case of the synchronization signal $r_H(n)$ being temporarily corrupt or unavailable. In addition, the use of a DPLL renders the previous point filtering and gap filling steps unnecessary.

There is a vast literature on the design and performance of both analog and digital

[2]To improve numerical stability and complexity one could also perform the LP analysis at a lower sample rate and then upsample the resulting error signal.

phase-locked loops [104, 105]. We start with a regular second-order all digital phase-locked loop [106]. Inspired by dual-loop gear-shifting DPLLs [107] we extended the DPLL to a dual-loop structure to achieve fast locking of the loop. We also dynamically adapt the bandwidth of the second loop in order to increase its robustness against spurious input signals.

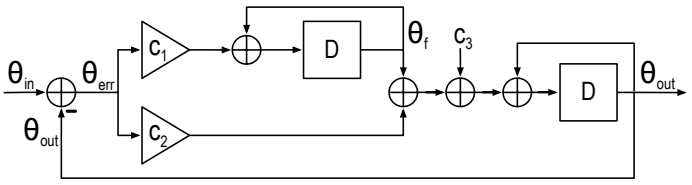


Figure 5.2.: Second-order all digital phase-locked loop.

The structure of the two loops is identical and is shown in Figure 5.2. The input signal $\theta_{in}(n)$ of the DPLL is the *phase angle* of the synchronization signal $r(n)$, which is given by the complex argument of the analytic signal of $r(n)$. The phase output $\theta_{out}(n)$ is converted back to a sinusoidal signal $r_d(n) = -\cos(\theta_{out}(n))$. The loop is computed and periodically updated with the following set of equations:

$$\theta_{in}(n) = \arg(r(n) + jr_H(n)) \qquad \theta_{err}(n) = \theta_{in}(n) - \theta_{out}(n)$$

$$\theta_f(n) = c_1\theta_{err}(n) + \theta_f(n-1)$$

$$\theta_{out}(n) = c_3 + \theta_f(n-1) + c_2\theta_{err}(n-1) + \theta_{out}(n-1)$$

The input and output phases $\theta_{in}(n)$ and $\theta_{out}(n)$ are always taken modulo-2π. The parameter $c_3 = \frac{2\pi f_s}{kf_s}$ specifies the initial frequency of the loop, with k being the oversampling factor of the synchronizer. The parameters c_1 and c_2 are related to the natural frequency $f_n = \frac{\omega_n}{2\pi}$ of the loop, which is a measure of the response time, and the damping factor $\eta = \frac{1}{\sqrt{2}}$, which is a measure of the overshoot and ringing, by

$$c_1 = \left(\frac{2\pi f_n}{kf_s}\right)^2 \qquad \text{and} \qquad c_2 = \frac{4\pi\eta f_n}{kf_s}.$$

The loop is stable if [106]

$$c_1 > 0, \qquad c_2 > c_1 \qquad \text{and} \qquad c_1 > 2c_2 - 4.$$

In the first loop we use a natural frequency $f_{n,1} = 5000\,\text{Hz}$, which enables a fast locking of the loop. However, the output $\theta_{out}(n)$ then closely resembles the input $\theta_{in}(n)$ of the loop, so that when the input is corrupted, also the output becomes corrupted. We consider the first loop as locked when the sample variance $\sigma_1^2(n)$ of the phase error $\theta_{err}(n)$ in a local window of ten samples is below a given threshold for a defined amount of time. We then gradually decrease the loop natural frequency $f_{n,2}$ of the second loop (which runs in parallel with the first loop) from $f_{n,2} = 5000\,\text{Hz}$ to $f_{n,2} = 50\,\text{Hz}$. This increases the loop's robustness against noise and jitter. When the

input signal is corrupt, the DPLL should continue running at the same frequency and not be affected by the missing or corrupt input. Therefore, when the variance of the phase error in the first loop increases, we reduce the natural frequency of the second loop to $f_{n,2}(n) = 10^{-\sigma_1^2(n)} 50\,\mathrm{Hz}$.

5.2.2. Data Frame Synchronization

This subsection describes the implementation and experimental validation of the frame synchronization method outlined in Section 5.1.3. To simplify the description, and because all experiments in Section 4.3 are based on a binary symbol alphabet, we use binary symbols herein, too.

5.2.2.1. Frame Grid Detection

To estimate the delay of the transmission channel and localize the frame grid position in the received signal, the received and sampled signal $s'(n)$ is correlated with the embedded training sequence m_{Tr}. The delay Δ between the embedder and the receiver can be estimated by the lag between the two signals where the absolute value of their cross-correlation is maximum, or

$$\Delta = \underset{k}{\operatorname{argmax}} \left|\rho_{s'm_{\mathrm{Tr}}}(k)\right| = \underset{k}{\operatorname{argmax}} \left| \sum_{n=-\infty}^{\infty} s'(n) m_{\mathrm{Tr}}(n-k) \right|. \tag{5.1}$$

The range of k in the maximization is application-dependent. It can be short if a defined utterance start point exists, which is the case in the aeronautical radio application. The sign of the cross-correlation at its absolute maximum also indicates if a phase inversion occurred during transmission. The delay is compensated and, in case a phase inversion occurred, the polarity adjusted, and

$$s'_{\mathrm{FB}}(n) = \operatorname{sgn}[\rho_{s'm_{\mathrm{Tr}}}(\Delta)] s'(n+\Delta).$$

5.2.2.2. Active Frame Detection

The frame synchronization scheme must be able to cope with detection errors in the frame preambles. A simple minimum Euclidean distance detector for the binary symbols in the equalized signal $e'(n)$ is the sgn function, and the detected binary symbol sequence m'_d of the d'th frame is

$$m'_d(k) = \operatorname{sgn}\left[e'(dL_D + k)\right]. \tag{5.2}$$

There are bit errors in m'_d due to the noise and the time-variant filtering of the channel.

To determine the active frames in the presence of bit errors, we denote with $n_{\mathrm{Ac}}(d)$ the number of samples in frame m'_d at the corresponding position that are equal to the marker sequence m_{Ac}. Equivalently, we denote with $n_{\mathrm{Tr}}(d)$ the number of samples that equal the training sequence m_{Tr}. We then consider those frames d''_a as *potentially* active frames where

$$\frac{n_{\mathrm{Ac}}(d)}{L_{\mathrm{Ac}}} \frac{n_{\mathrm{Tr}}(d)}{L_{\mathrm{Tr}}} \geq \tau_S \tag{5.3}$$

with τ_S being a decision threshold (with $0 \leq \tau_S \leq 1$). In the case of zero bit errors, the product in (5.3) would evaluate to unity in active frames, and to 0.25 on average in non-active frames. We use a decision threshold $\tau_S = 0.6$, which was determined experimentally.

Insertions, deletions and substitutions (IDS) occur in the detection of the active frame candidates d''_a. To accommodate for this, we incorporate the frame indices m_{ID}, which sequentially number the active frames and are encoded so as to maximize the Hamming distance between neighboring frame indices. To improve the detection accuracy of the frame indices m_{ID}, the signals $r'(n)$ and m'_d are recomputed using the active frame markers m_{Ac} in the frame candidates d''_a as additional training symbols for equalization.

The sequence of decoded frame indices Y' detected in the frames d''_a is not identical to the original sequence of indices $\mathrm{Y} = 1, 2, 3, \ldots$, because of the IDS errors in d''_a and the bit errors in m'_d. We use a dynamic programming algorithm for computing the Levenshtein distance between the sequences Y and Y', in order to determine where the sequences match, and where insertions, deletions and substitutions occurred [108]. We process the result in the following manner:

1. Matching frame indices are immediately assumed to be correct active frames.

2. In the case of an insertion in Y', either the insertion or an adjacent substitution in Y' must be deleted. The one active frame candidate is removed from the sequence that minimizes as error measure the sum of the number of bit errors in the frame indices of the remaining frame candidates.

3. In the case of a deletion in Y', first all substitutions surrounding the deletion are removed, too. Then, the resulting 'hole' of missing active frames must be filled with a set of new active frames. Out of all possible frames that lie between the adjacent matching frames, a fixed number of frames (given by the size of the hole) needs to be selected as active frames. Out of all possible frame combinations the one set is selected that minimizes as error measure the sum of the number of bit errors in the frame identification, the active frame marker and the training sequence, over all frames in the candidate set. To give higher weight to consecutive frames, we reduce the error distance by one if a frame candidate has a direct neighbor candidate or is adjacent to a previously matched active frame.

Using the above detection scheme, and assuming the total number of frames embedded in the received signal to be known, there are by definition no more insertion or deletion errors, but only substitution errors in the set of detected active frames d'_a. This then also applies for the embedded watermark message, which considerably simplifies the error correction coding of the payload data since a standard block or convolution code can be applied.

5.3. Experimental Results and Discussion

5.3.1. Timing Recovery and Bit Synchronization

5.3.1.1. Non-Data-Aided Bit Synchronization

Experimental Setup The spectral line based timing recovery algorithm presented in this chapter is necessary only when using an LP-based watermark detector. LP-based watermark detection is only possible when the transmission channel has no filtering distortion. Since in this case many components of the watermark embedding scheme presented in Chapter 4 can be omitted, we evaluate the timing recovery using the simpler watermark embedding system that we presented in [2]. The system omits the bandpass embedding and pulse shaping parts, assumes ideal frame synchronization, and performs LP-based watermark detection.

An unsynchronized resampling of the received signal leads to a timing phase error, which is a phase shift of a fractional sample in the sampling instants of the watermark embedder and detector.

We use a piecewise cubic Hermite interpolating polynomial (PCHIP) to simulate a reconstruction of a continuous-time signal and resample the resulting piecewise polynomial structure at equidistant sampling points at intervals of $\frac{1}{f_s}$ or $\frac{1}{kf_s}$ respectively. In this reconstruction process, each sample of the watermarked speech signal $\hat{s}(n)$ serves as a data point for the interpolation. The nodes of these data points would ideally reside on an evenly spaced grid with intervals of $\frac{1}{f_s}$. In order to simulate an unsynchronized system we move the nodes of the data points to different positions according to three parameters:

Timing phase offset: All nodes are shifted by a fraction of the regular grid interval $\frac{1}{f_s}$ (unit interval, UI).

Sampling frequency offset: The distance between all nodes is changed from one unit interval to a slightly different value.

Jitter: The position of each single node is shifted randomly following a Gaussian distribution with variance σ_J^2.

The experiments presented in this subsection with the LP-based watermark detector use as input signal a short sequence of noisy air traffic control radio speech with a length of 6 s and a sampling frequency of $f_s = 8000$ Hz.

Results Without synchronization, the watermarking system is inherently vulnerable to an offset of the sampling phase and frequency between the embedder and the detector. Figure 5.3 shows the adverse effect of a timing phase offset if no synchronization system is used. We measure this timing phase error in 'unit intervals' (UI), which is the fraction of a sampling interval $T = \frac{1}{f_s}$.

The proposed synchronization system aims to estimate the original position of the sample nodes, which is the optimal position for resampling the continuous-time signal

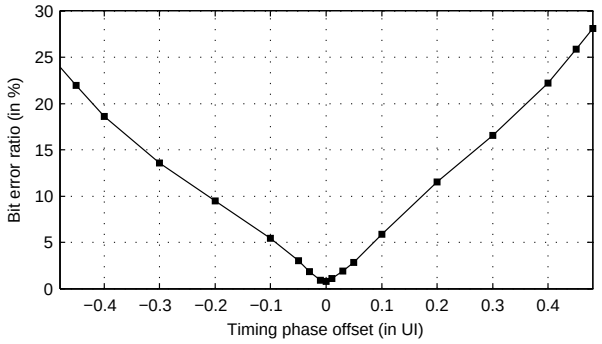


Figure 5.3.: Robustness of the LP-based watermark detector without synchronization with respect to a timing phase error.

at the detector. The phase estimation error is the distance between the original and the estimated node position in unit intervals. Its root-mean-square value across the entire signal is shown in Figure 5.4 for the above three types of synchronization errors. The figure also shows the bit error ratio for different sampling frequency offsets. The bit error ratio as a function of the uncorrected timing phase offset is shown in Figure 5.3.

Figure 5.5 shows the raw bit error ratio of the overall watermarking system (in two different configurations and including a watermark floor and residual emphasis as described in [2]) at different channel SNR and compares the proposed synchronizer to artificial error-free synchronization. Compared to the case where ideal synchronization is assumed, the raw BER increases by less than two percentage points across all SNR values.

5.3.1.2. Data-Aided Bit Synchronization

We experimentally evaluate the capability of the RLS equalization based watermarking system (Chapter 4) to compensate a timing synchronization offset. In contrast to the previous experiments, we use the system implementation and experimental settings of Section 4.3, again using ten randomly selected utterances of the ATCOSIM corpus as input speech. Figure 5.6(a) shows the resulting BER when introducing a timing phase offset of a fractional sample in the simulated channel as described above.

The error ratio significantly increases for timing phase offsets around -0.4 UI (equivalent to an offset of 0.6 UI). This results from the fact that the RLS equalizer is a causal system and cannot shift the received signal forward in time. The increased error rate can be mitigated by delaying the training sequence that is fed to the equalizer by one sample in time and adding one equalizer filter tap ($K = 12$). As shown in Figure 5.6(b), the BER is then below 8 % for any timing phase offset.

The remaining increase in BER for non-zero timing phase offsets is a result of aliasing. To overcome this limitation, a fractionally spaced equalizer should be used instead of the sample-spaced equalizer of Chapter 4 [109, 110]. The discontinuity in the BER curve at 0.6 UI results from the shift of the detected frame grid towards the next sample (cf. Section 5.2.2.1).

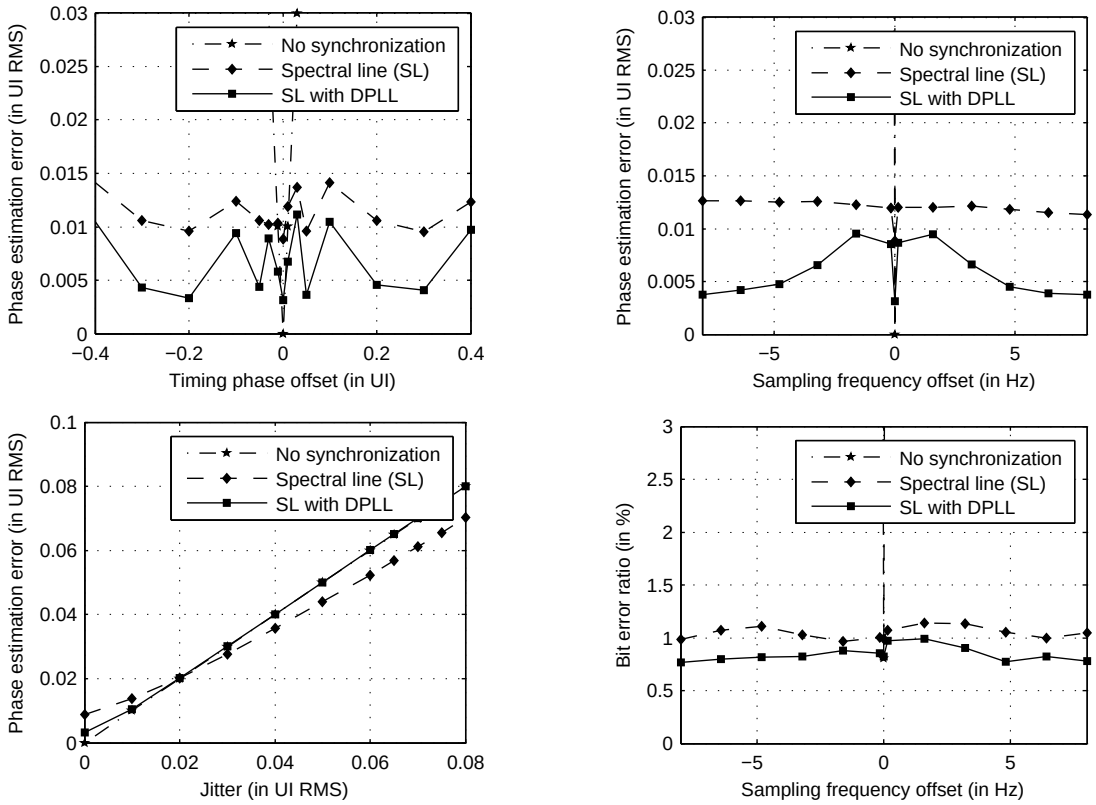


Figure 5.4.: Synchronization system performance for LP-based watermark detection: Phase estimation error and bit error ratio for various types of node offsets.

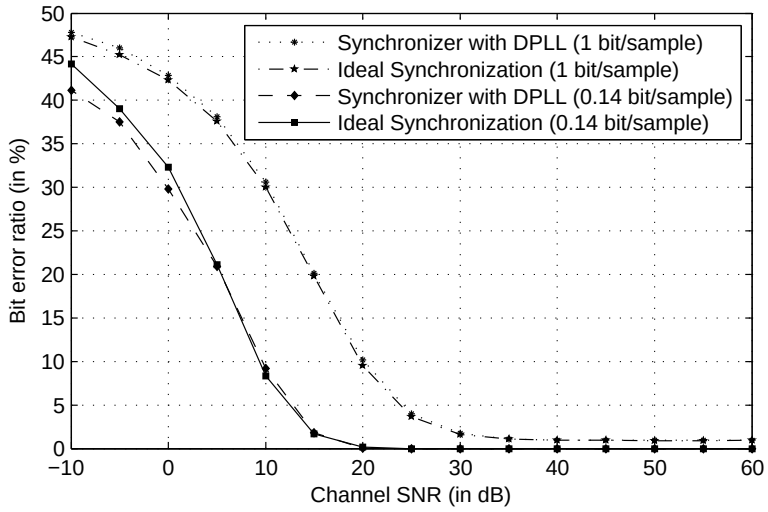


Figure 5.5.: Comparison of overall system performance (using the LP-based watermark detector) with ideal (assumed) synchronization and with the proposed synchronizer at two different embedding rates (in watermark bits per host signal sample).

5.3.2. Data Frame Synchronization

We evaluate the data frame synchronization in the context of the RLS equalization based watermarking system discussed in Chapter 4.

5.3.2.1. Frame Grid Detection

In certain applications a near-real-time availability of the watermark data is required. To evaluate the time needed until frame grid synchronization is achieved, in the following experiment we used only the first $T_{\mathrm{FG}} = 0\,\mathrm{s} \ldots 4\,\mathrm{s}$ of each utterance to detect the frame grid. Figure 5.7 shows the fraction of utterances with correct frame grid detection for a given signal segment length T_{FG}, evaluated over 200 randomly selected utterances of the ATCOSIM corpus and a clean transmission channel. The mean fraction of non-voiced samples in the corresponding segments (also shown in Figure 5.7) explains the decrease of the detection ratio for $T_{\mathrm{FG}} > 0.2\,\mathrm{s}$.

The experiment shows that a correct frame grid detection can be achieved for correlation windows as short as $T_{\mathrm{FG}} = 50\,\mathrm{ms}$ (Figure 5.7). The detection ratio locally decreases for longer correlation windows because most utterances in the used database start with a non-voiced segment. Because m_{Tr} is embedded in the non-voiced segments only, correlation is low when the window is still relatively short and the fraction of non-voiced samples within the window is low. This effect could be avoided by considering only the non-voiced segments in the correlation sum (5.1), which, however, requires a voiced/non-voiced segmentation in the detector.

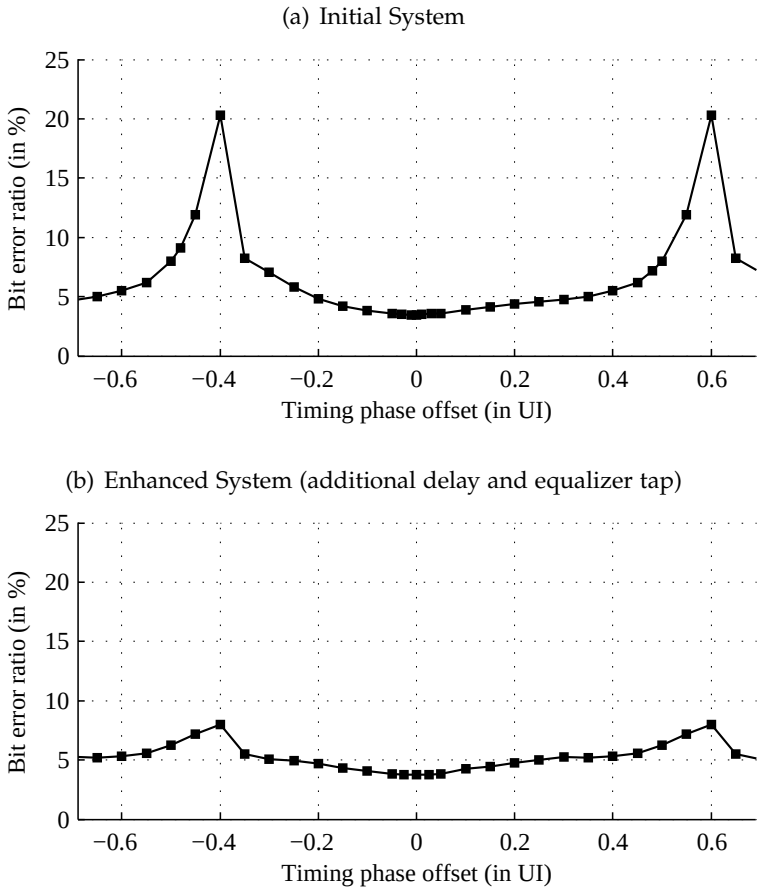


Figure 5.6.: Overall system robustness in the presence of a timing phase offset at an average uncoded bit rate of 690 bit/s.

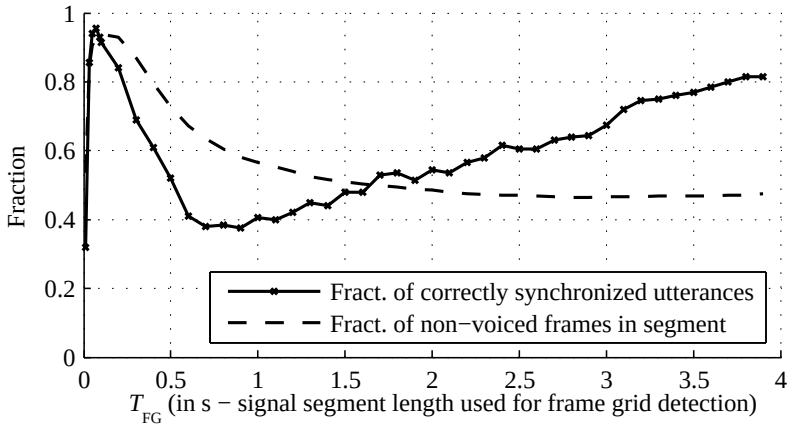


Figure 5.7.: Fraction of successfully synchronized utterances as a function of the signal segment length used for the frame grid detection.

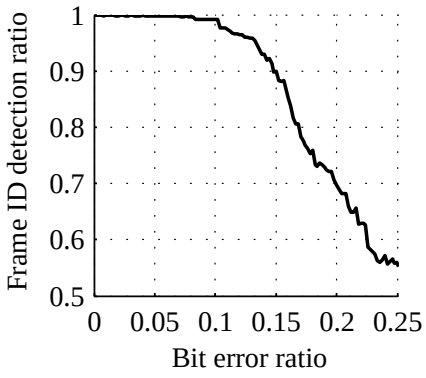


Figure 5.8.: Robustness of active frame detection against bit errors in the received binary sequence m'_d.

5.3.2.2. Active Frame Detection

In order to evaluate the performance of the active frame detection subsystem, we introduced artificial bit errors into an error-free version of the detected binary sequence m'_d of (5.2), resulting from ten randomly selected utterances of the ATCOSIM corpus. Figure 5.8 shows the ratio of correctly detected frames as a function of the bit error ratio (BER) in m'_d. The fraction of spurious frames is one minus the given ratio.

The experiment shows that the active frame detection is highly reliable up to a bit error ratio of 10 % in the detected binary sequence. It was further observed that active frame detection errors only contribute to a small extent to the overall system bit error ratio: Using ideal frame synchronization instead of the proposed scheme, the change in the bit error ratios of Figure 4.7 is < 0.5 %-points for all channels.

5.4. Conclusions

We conclude that timing recovery and bit synchronization can be achieved without additional effort using the RLS equalization based watermark detector of Chapter 4. In addition, the RLS equalizer resolves many problems occurring otherwise in the synthesis/analysis system synchronization.

We demonstrated the application of classical spectral line synchronization in the context of watermarking for a simpler but less robust linear prediction based watermark detector. In combination with a digital PLL, the synchronization method provides highly accurate timing recovery.

Frame grid synchronization can be achieved with windows as short as 50 ms, but leaves room for further improvements for example by incorporating a voiced/non-voiced detection in the detector. On the contrary, the active frame detection works reliably up to a bit error ratio of 10 % and does not significantly contribute to the overall watermark detection errors.

Part II.

Air Traffic Control

Chapter 6

Speech Watermarking for Air Traffic Control

The aim of civil air traffic control (ATC) is to maintain a safe separation between all aircraft in the air in order to avoid collisions, and to maximize the number of aircraft that can fly at the same time. Besides a set of fixed flight rules and a number of navigational systems, air traffic control relies on human air traffic control operators (ATCO, or controllers). The controller monitors air traffic within a so-called sector (a geographic region or airspace volume) based on previously submitted flight plans and continuously updated radar pictures, and gives flight instructions to the aircraft pilots in order to maintain a safe separation between the aircraft.

Although digital data communication links between controllers and aircraft are slowly emerging, most of the communication between controllers and pilots is verbal and by means of analog voice radios. Air traffic control (ATC) has relied on the voice radio for communication between aircraft pilots and air traffic control operators since its beginning. The amplitude-modulation (AM) radio, which is in operation worldwide, has basically remained unchanged for decades. The AM radio is based on the double-sideband amplitude modulation (DSB-AM) of a sinusoidal carrier. For the continental air-ground communication, the carrier frequency is within a range from 118 MHz to 137 MHz, the 'very high frequency' (VHF) band, with a channel spacing of 8.33 kHz or 25 kHz [97]. Given the aeronautical life cycle constraints, it is expected that the analog radio will remain in use for ATC voice in Europe well beyond 2020 [111, 112].

Parts of this chapter have been published in K. Hofbauer and H. Hering, "Digital signatures for the analogue radio," in *Proceedings of the NASA Integrated Communications Navigation and Surveillance Conference (ICNS)*, Fairfax, VA, USA, 2005.

6.1. Problem Statement

The avionic radio is the main tool of the controller for giving flight instructions and clearances to the aircraft pilot. It is crucial for a secure and safe air traffic operation to have a reliable and fail-safe radio communication network to guarantee the possibility of communication at any given time. From a technical point of view, high effort is put into the systems in order to provide permanent availability through robust and redundant design.

Once this 'technical' link between ground and aircraft is established (which we assume further-on), the verbal communication between pilot and controller can start. In order to avoid misunderstandings and to guarantee a common terminology, the two parties use a restricted, very simple, common language. Although most of the words are borrowed from the English language, the terms and structure of this language are clearly defined in the corresponding standards [113]. Every voice communication on the aeronautical channel is supposed to take place according to this ICAO terminology.

The radio communication occurs on a party-line channel, which means that all aircraft within a sector as well as the responsible controller transmit and listen to all messages on the corresponding radio channel frequency. In order to establish a meaningful communication in this environment, it has to be clear who is talking (the addresser, sender, originator) and to whom the current word is addressed to (the addressee, recipient, acceptor). For the ATC air-ground communication, certain rules are in place to establish this identification. In the standard situations, the air traffic controller is the only person on the channel that does not identify himself as addresser on the beginning of the message. Instead, the controller starts the message with the call-sign of the aircraft, the addressee of the message. In case not otherwise explicitly specified, every voice message of the aircraft pilot is inherently addressed to the air traffic controller. Therefore the identification of the addressee (the controller in this case) is omitted, and the pilot starts the message with the flight's call-sign to identify the addresser of the message.

The correct identification of addresser and addressee is crucial for a safe communication, and there are a number of problems associated with this identification process:

1. *Channel monitoring workload.* The aircrew needs to carefully monitor the radio channel to identify messages which are addressed to their aircraft. This creates significant workload.

2. *Aircraft identification workload.* The controller needs to identify the aircraft first in the radio call, then on the radar screen. There is no convergence between the radio call and the radar display system, which creates additional controller workload.

3. *Wrongly understood identification.* An aircraft call-sign can be wrongly understood by the controller or the pilot. This highly compromises safety, and is most likely to happen when aircraft with similar call-sign are present in the same sector. This potential risk is usually referred to as call-sign confusion.

4. *Missed identification.* If the identification is not understood by the addressee, the entire message is declared void and has to be repeated. This is additional workload for both the aircrew and the controller.

5. *Forged identification.* There is currently no technical barrier preventing a third party to transmit voice messages with a forged identification. This security threat is currently only addressed through operational procedures and the party-line principle of the radio channel.

An automatic identification and authentication of the addresser and the addressee could solve these issues and, thus, improve the safety and security in air traffic control.

6.2. High-Level System Requirements

Developing an application for the aviation domain implies some special peculiarities and puts several constraints on the intended system. From the above scenario we can derive a list of requirements which such an authentication system has to fulfill. The user-driven requirements specify the functions and features that are necessary for the users of the system. At the same time various aspects have to be considered to enable the deployment of the system within the current framework of avionic communications.

6.2.1. Deployment-Driven Requirements

Rapid Deployment First and foremost the development should keep a rapid implementation and a short time to operational use in mind. The basic approach is to improve the current communication system, as call-sign confusion and ambiguity is currently an unresolved issue and will become even more problematic as air traffic increases. Even though a long-term solution for avionic air-ground communication will probably lie outside the scope of analog voice radio, analog radio will continue to be used worldwide for many years to come.

Legacy System Compliance The system should be backward compatible to the legacy avionic radio system currently in use. Changing to a completely new radio technology would have many benefits. However, it is very difficult to obtain a seamless transition to a new system. Neither is it possible to change all aircraft and ground equipment from one day to another, nor is it trivial to enforce deployment of new technologies among all participants. Co-existence among old and new systems is necessary and they should ideally complement each other.

Bandwidth Efficiency Ideally the system should not create any additional demands on frequency bandwidth. Especially in Europe there is a severe lack of frequency bandwidth in the preferable VHF band, with the future demand only increasing. This led to the reduction of the frequency spacing between two adjacent VHF channels from 25 kHz to 8.33 kHz, but this being a temporary solution. That is, a system that operates within the existing communication channels would be highly preferable.

Minimal Aircraft Modifications Changes to the certified aircraft equipment should be avoided as much as possible. Every new or modified piece of equipment, might it be a different radio transceiver or a connection to the aircraft data bus, entails a long and costly certification process. Through minimizing the changes to on-board equipment and maintaining low complexity, the certification process can be highly simplified.

Cost Efficiency Finally, the total costs of necessary on-board equipment should be kept as low as possible. Not only that there are much higher numbers of airborne than ground systems, but as well most of the ground radio stations are operated by the air traffic authorities, which are generally less cost-sensitive than airlines. Therefore the development process should, wherever possible, shift costs from on-board to ground systems.

6.2.2. User-Driven Requirements

Perceptual Quality Several potential identification systems affect the perceptual quality of the speech transmission. A too severe degradation of the sound quality would not be accepted by both the certification authorities and the intended users, and for example become annoying to the air traffic controllers. Ideally the participants of the communication should not notice the presence of the system. For this reason the perceptual distortion needs to be kept at a minimum.

Real-Time Availability The identification of the addresser has to be detected immediately at the beginning of the voice message. If the identification would display only after the voice message, this would be of little use to the air traffic controller, as he is then already occupied with another task. The automatic identification should be completed before the call-sign is completely enunciated, ideally in less than one second.

Data Rate In order to serve the purpose of providing identification of the addresser, it is necessary to transmit a certain amount of data. This might be the aircraft's tail number or the 27 bit Data Link Service Address for aircraft and ground stations used in the context of Controller Pilot Data Link Communication (CPDLC). A GPS position report, which would be advantageous in various scenarios, requires approximately 50 bit of payload data. Altogether a data rate of 100 bit/s is desired, leaving some room for potential extensions. For a secure authentication, most likely a much higher data rate is required.

Error Rate Robustness is a basic requirement. An unreliable system would not be accepted by the users. Two types of errors are possible. On the one hand, the system can fail to announce the addresser altogether, and although this not safety critical, too frequent occurrence would lead to user's frustration. On the other hand

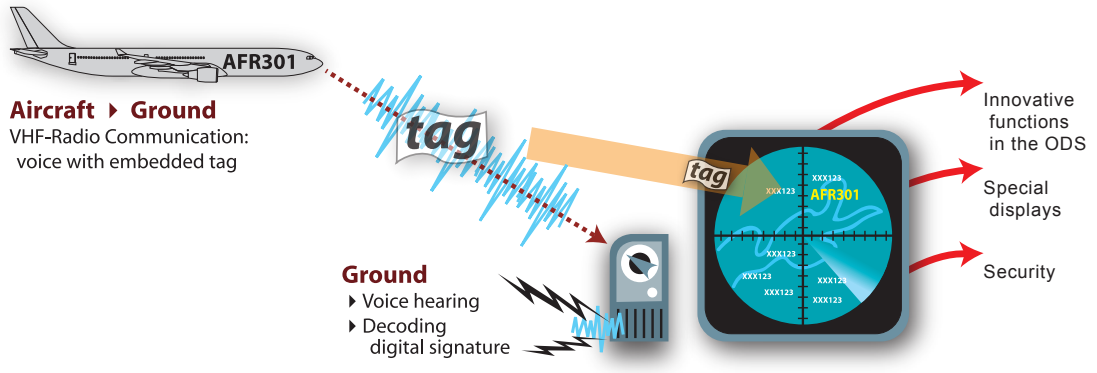


Figure 6.1.: Identification of the transmitting aircraft through an embedded watermark.

the announcement of a wrong addresser can compromise safety. Therefore, it is indispensable to assure robust transmission and to verify the data's validity.

Maintaining Established Procedures Due to the strong focus on safety, the organization and work-flow in commercial aviation and air traffic control is highly based on procedures, of which some are in place for decades. As a consequence, whenever changes to the procedures are proposed, there is a thorough and time-demanding review and evaluation process. For a rapid deployment it therefore seems beneficial not to change or replace these procedures, but to provide supplementary services, only.

No User Interaction The system should be autonomous and transparent to the user. For the controller, the system should support his work by providing additional information, and should not add another task. For the pilots, a need for additional training and additional workload is unwanted. Therefore, the system should work without any human intervention.

6.3. Background and Related Work

6.3.1. Inband Data Transmission Using Watermarking

In order to fulfill the above requirements for an automatic identification system, the use of speech watermarking was previously proposed [114, 28]. Figure 6.1 shows the overall idea.

The general components of a speech watermarking system for the aeronautical application are shown in Figure 6.2. The voice signal is the host medium that carries the watermark and is produced from the speaker's microphone or headset. The watermark, the data that is embedded into the voice signal, could for example consist of the 24 bit aircraft identifier and—depending on the available data rate—auxiliary data such as the aircraft's position or an authentication. The watermark embedder could be fitted into a small adapter box between the headset and the existing VHF

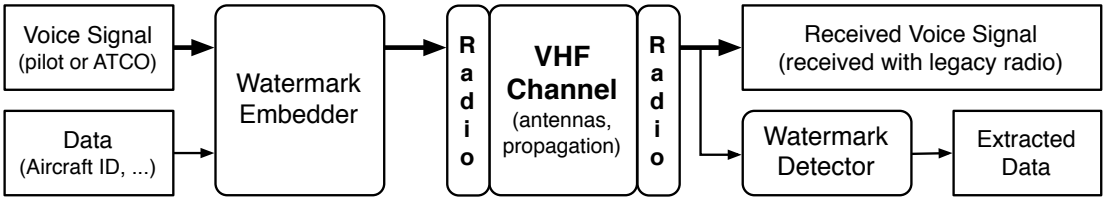


Figure 6.2.: General structure of a speech watermarking system for the aeronautical voice radio.

radio [115]. It converts the analog speech signal to the digital domain, embeds the watermark data, and converts the watermarked digital signal back to an analog speech signal for transmission with the standard VHF radio.

The transmission channel consists of the airborne and ground-based radio transceivers, corresponding wiring, antenna systems, etc., and the VHF radio propagation channel. The channel has crucial influence on the performance of the system and is therefore the subject of investigation in subsequent chapters. Although the transmitted signal contains a watermark, it is technically and perceptually very similar to the original speech signal and can therefore be received and listened to with every standard VHF radio receiver without any modifications. This allows a stepwise deployment and parallel use with the current legacy system. The watermark detector extracts the data from the received signal, assures the validity of the data and displays it to the user. The information could also be integrated into the ATC systems, e.g., by highlighting the radar screen label of the aircraft that is currently transmitting. An industrial study produced a detailed overview on operational concepts, potential benefits, applications and packages, equipment and implementation scenarios, as well as a cost-benefit analysis [116].

6.3.2. Related Work

This thesis focuses on the algorithms for the watermark embedding and detection, and the aeronautical radio channel. Figure 6.3 shows how the area of research narrows down by considering the constraints as outlined in Section 6.2.

Transmitting the identification tag is first and foremost a data communications problem. As the transmission should occur simultaneously within the legacy voice channel, a speech watermarking system has to be applied. As shown in Section 2.3, the body of research concerning watermarking tailored to speech signals is small. While most of the existing watermarking schemes are tailored to be robust against lossy perceptual coding, this requirement does not apply to watermarking for the aeronautical radio communication. This thesis ultimately focuses on speech watermarking for the noisy and time-variant analog aeronautical radio channel, and we give in the following a short summary on the few existing systems tailored to the aeronautical application.

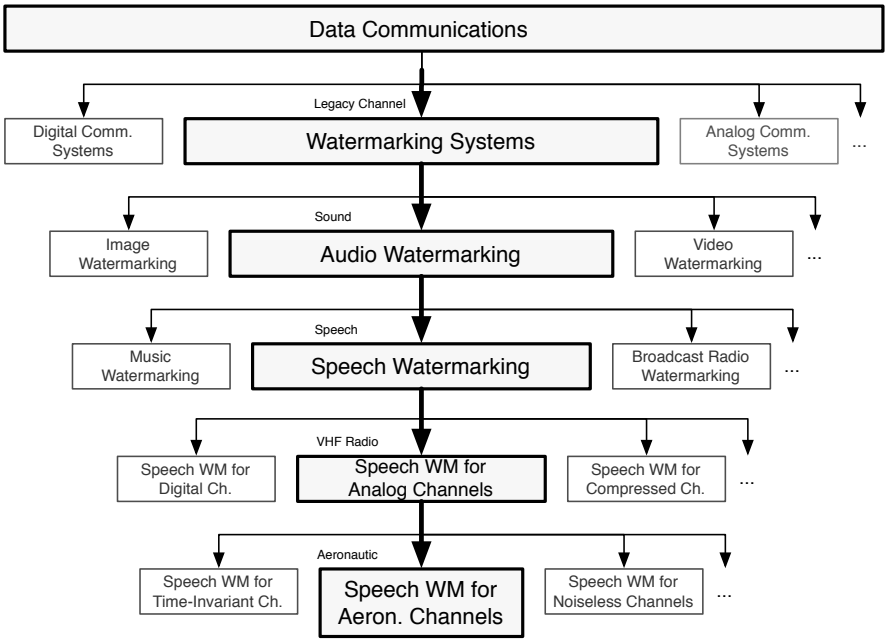


Figure 6.3.: Research focus within the field of data communications.

6.3.2.1. Multitone Sequence Transmission

The most elementary technique to transmit the sender's digital identification is a multitone data sequence at the beginning of the transmission [117]. Standard analog modulation and demodulation schemes can be applied to create a short high power data package which is transmitted before each message. This is a very simple and field-proven technology, which provides high robustness and data rates. The transmission is clearly audible to the listener as a noise burst at the beginning of the message. The noise burst, which would hardly be accepted by the user base, can be removed by an adaptive filter. However, the fact that all of the receivers would have to be equipped with this technology renders the entire system undesirable.

6.3.2.2. Data-in-Voice (DiV) Modem

The 'Data-in-Voice' system presented in [118] tries to decrease the audibility of the multitone sequence by reducing its power and spectral bandwidth. Using a band-reject filter, the system removes some parts of the voice spectrum around 2 kHz as to operate an in-band modem. A standard 240 bit/s MSK (Minimum Shift Keying) modem technique is used, which occupies a frequency bandwidth of approximately 300 Hz. In order to fully suppress the audibility of the data sequence, the system also requires a filter on the side of the voice receivers.

6.3.2.3. Spread Spectrum Watermarking

The 'Aircraft Identification Tag' (AIT) system presented in [28, 119] is based on direct sequence spread spectrum watermarking and embeds the watermark as additive pseudo-random white noise. The embedder first adds redundancy to the data by an error control coding scheme. The coded data is spread over the available frequency bandwidth by a well-defined pseudo-noise sequence. The watermark signal is then spectrally shaped with a linear predictive coding filter and additively embedded into the digitized speech signal, thus exploiting frequency masking properties of human perception. The detector relies on a whitening filter to compensate for the spectral shaping of the watermark signal produced by the embedder. A maximum-length pseudo random sequence (ML-PRS), detected with a matched filter, is used to ensure synchronization between embedder and detector. The signal is then de-spread and the watermark data extracted. The voice signal acts as additive interference (additive noise) that deteriorates the detector's ability to estimate the watermark. Therefore, even in the case of an ideal noiseless transmission channel, the data rate is inherently limited.

6.3.3. Proposed Solution

For the embedding of an identification or other data in the ATC radio speech, we propose to use the speech watermarking algorithm presented in Chapter 4, or a variant thereof. To substantiate the practical feasibility of our approach, a thorough evaluation and a large number of further tests is required, and we focus in the following chapters on two important aspects.

First, the embedding capacity of the proposed algorithm is host speech signal dependent. In order to obtain realistic simulation results given the particular ATC phraseology and speaking style, we produced and present in Chapter 7 a speech database of ATC operator speech. The resulting ATCOSIM corpus is not only used for evaluating the watermarking method, but is also of general value for ATC-related spoken language technologies.

Second, the aeronautical transmission channel has a crucial influence on the data capacity and robustness of the overall system. We therefore study the aeronautical voice channel properties both in terms of stochastic channel models as well as using empirical end-to-end measurements and derive a simulation model based on the measurements.

Given the speech corpus and the results of the channel analysis, Chapter 10 will validate the suitability of the proposed watermarking scheme for the aeronautical application.

Chapter 7

ATC Simulation Speech Corpus

The ATCOSIM Air Traffic Control Simulation Speech corpus is a speech database of air traffic control (ATC) operator speech. ATCOSIM is a contribution to ATC-related speech corpora: It consists of ten hours of speech data, which were recorded during ATC real-time simulations. The database includes orthographic transcriptions and additional information on speakers and recording sessions. The corpus is publicly available and provided free of charge. Possible applications of the corpus are, among others, ATC language studies, speech recognition and speaker identification, the design of ATC speech transmission systems, as well as listening tests within the ATC domain.

Parts of this chapter have been published in K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM corpus of non-prompted clean air traffic control speech," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008. More detailed information is available in [120].

7.1. Introduction

Until today spoken language technologies such as automatic speech recognition are close to non-existing in operational air traffic control (ATC). This is in parts due to the high reliability requirements that are naturally present in air traffic control. The constant progress in the development of spoken language technologies opens a door to the use of such techniques for certain applications in the air traffic control domain. This is particularly the case for the controller speech on ground, considering the good signal quality (close-talk microphone, low background noise, known speaker) and the restricted vocabulary and grammar in use. In contrast, doing for example speech recognition for the incoming noisy and narrowband radio speech is still a quite difficult task.

In the development of practical systems the need for appropriate corpora comes into place. The quality of air traffic control speech is quite particular and falls in-between the classical categories: It is neither spontaneous speech due to the given constraints, nor is it read, nor is it a pure command and control speech (in the sense of controlling a device). Due to this and also due to the particular pronunciation and vocabulary in air traffic control, there is a need for speech corpora that are specific to air traffic control. This is even more the case considering the high accuracy and robustness requirements in most air traffic control applications.

The international standard language for ATC communication is English. The use of the French, Spanish or Russian language is also permitted if it is the native language of both pilot and controller involved in the communication. The phraseology that is used for this communication is strictly formalized by the International Civil Aviation Organization [113]. It mandates the use of certain keywords and expressions for certain types of instructions, gives clear rules on how to form digit sequences, and even defines non-standard pronunciations for certain words in order to account for the band-limited transmission channel. In practice however, both controllers and pilots deviate from this standard phraseology.

After a review of the existing ATC related corpora known to the authors, the subsequent sections present the new ATCOSIM Air Traffic Control Simulation Speech corpus, which fills a gap that is left by the existing corpora. Section 7.2 outlines the difficulty of obtaining realistic air traffic control speech recordings and shows the path chosen for the ATCOSIM corpus. The transcription and production process is described in Section 7.3 and 7.4. We conclude with a proposal for a specific automatic speech recognition (ASR) application in air traffic control.

ATC Related Corpora

Despite the large number of existing speech corpora, only a few corpora are in the air traffic control domain.

The *NIST Air Traffic Control Complete Corpus* [121] consists of recordings of 70 hours of approach control radio transmissions at three airports in the United States. The recordings are narrowband and of typical AM radio quality. The corpus contains an orthographic transcription and for each transmission the corresponding flight number

is listed. The corpus was produced in 1994 and is commercially available.

The *HIWIRE* database [122] is a collection of read or prompted words and sentences taken from the area of military air traffic control. The recordings were made in a studio setting, and cockpit noise was artificially added afterwards. The database contains 8,100 English utterances pronounced by non-native speakers without air traffic control experience. The corpus is available on request.

The *non-native Military Air Traffic Control (nnMATC)* database [123] is a collection of 24 hours of military ATC radio speech. The recordings were made in a military air traffic control center, wire-tapping the actual radio communication during military exercises. The recordings are narrowband and of varying quality depending on the speaker location (control room or aircraft). The database was published in 2007, but its use is restricted to the NATO/RTO/IST-031 working group and its affiliates.

The *VOCALISE* project [124, 125] recorded and analyzed 150 hours of operational ATC voice radio communication in France, including en route, approach and tower control. The database is not available for the public and its use is restricted to research groups affiliated with the French 'Centre d'Études de la Navigation Aérienne' (CENA)—now part of the 'Direction des Services de la Navigation Aérienne' (DSNA).

Another corpus that is similar to ours, but outside the domain of ATC, is the *NATO Native and Non Native (N4) database* [126]. It consists of ten hours of military naval radio communication speech recorded during training sessions.

The aforementioned corpora vary significantly among each other with respect to e.g. scope, technical conditions and public availability (Table 7.1). The aim of the ATCOSIM corpus is to fill the gap that is left by the above corpora: ATCOSIM provides 50 hours of publicly available direct-microphone recordings of operational air traffic controller speech in a realistic civil en-route control situation. The corpus includes an utterance segmentation and an orthographic transcription. ATCOSIM is meant to be versatile and is as such not tailored to any specific application.

7.2. ATCOSIM Recording and Processing

The aim of the ATCOSIM corpus production was to provide wideband ATC speech which should be as realistic as possible in terms of speaking style, language use, background noise, stress levels, etc.

In most air traffic control centers the controller pilot radio communication is recorded and archived for legal reasons. However, these legal recordings are problematic for a corpus production for a multitude of reasons. First, most recordings are based on a received radio signal and thus not wideband. Second, it is in general difficult to get access to these recordings. And third, even if one would obtain the recordings, their public distribution would be legally problematic at least in many European countries.

The logical resort is the conduction of simulations in order to generate the speech samples. However, the required effort to set up realistic ATC simulations (including facilities, hard- and software, trained personal, ...) is large and would far exceed the budget of a typical corpora production. However, such simulations are performed for the sake of evaluating air traffic control and air traffic management concepts, also on a

Table 7.1.: Feature Comparison of Existing ATC-Related Speech Corpora and the ATCOSIM Corpus

	NIST	**HIWIRE**	**nnMATC**	**VOCALISE**	**ATCOSIM**
Recording Situation					
- Recording content	civil ATCO & PLT	N/A	military ATCO & PLT	civil ATCO & PLT	**civil ATCO**
- Control position	approach	N/A	military	mixed	**en-route**
- Geographic region	USA	N/A	Europe (BE)	Europe (FR)	**Europe (DE/CH/FR)**
- Speaking style (context)	operational	prompted text	operational	operational	**operational** [(1)]
Recording Setup					
- Speech bandwidth	narrowband	wideband	mostly narrowband	unknown	**wideband**
- Transmission channel	radio	none	none / radio	none / radio	**none**
- Radio transmission noise	high	none	mixed	mixed	**none**
- Acoustical noise	CO & CR	CO (artificial)	CO & CR	CO & CR	**CR**
- Signal source	VHF radio	direct microphone	mixed	unknown	**direct microphone**
- Duration [active speech]	70 h [?]	(8100 utterances)	700 h [20 h]	150 h [45h]	**51.4 h [10.7 h]**
Speaker Properties					
- Level of English	mostly native	non-native	mostly non-native	mixed	**non-native**
- Gender	mixed	mixed	mostly male	mixed	**mixed**
- Operational	yes	no (!)	yes	yes	**yes**
- Field of prof. operation	civil	N/A	military	civil	**civil**
- Number of speakers	unknown (large)	81	unknown (large)	unknown (large)	**10**
Publicly Available	yes	yes (?)	no	no	**yes**

[(1)] Large-scale real-time simulation

• CO: Cockpit • CR: Control Room • ATCO: Controller • PLT: Pilot

large scale involving tens of controllers. The ATCOSIM speech recordings were made at such a facility during an ongoing simulation.

7.2.1. Recording Situation

The voice recordings were made in the air traffic control room of the EUROCONTROL Experimental Centre (EEC) in Brétigny-sur-Orge, France (Figure 7.1).[1] The room and its controller working positions closely resemble an operational control center room. The simulations aim to provide realistic air traffic scenarios and working conditions for the air traffic controller. Several controllers operate at the same time, in order to simulate also the inter-dependencies between different control sectors. The controller communicates via a headset with pseudo-pilots which are located in a different room and control the simulated aircraft. During the simulations only the controllers' voice, but not the pilots', was recorded, because the working environment of the pseudo-pilots, and as such the speaking style, did not to any extent resemble reality.

7.2.2. Speakers

The participating controllers were all actively employed air traffic controllers and possessed professional experience in the simulated sectors. The six male and four female controllers were of either German or Swiss nationality and had German, Swiss German or Swiss French native language. The controllers had agreed to the recording of their voice for the purpose of language analysis as well as for research and development in speech technologies, and were asked to show their normal working behavior.

7.2.3. Recording Setup

The controller's speech was picked up by the microphone of a Sennheiser HME 45-KA headset. The microphone signal and a push-to-talk (PTT) switch status signal were recorded onto digital audio tape (DAT) with a sampling frequency of 32 kHz and a resolution of 12 bit. The push-to-talk switch is the push-button that the controller has to press and hold in order to transmit the voice signal on the real-world radio. The speech signal was automatically muted when the push-button was not pressed. This results in a truncation of the speech signal if the push-button was pressed too late or released too early. Figure 7.2 shows an example of the recorded signals. The recorded PTT signal is a high-frequency tone that is active when the PTT switch is not pressed. After the digital transfer of the DAT tapes onto a personal computer and an automatic format conversion to a resolution of 16 bit by setting the less significant bits to zero, the status signal of the push-to-talk button could after some basic processing be used to reliably perform an automatic segmentation of the recorded voice signal into separate controller utterances.

[1] The raw recordings were collected by Horst Hering (EUROCONTROL) for a different purpose and prior to the work presented herein.

Figure 7.1.: Control room and controller working position at the EUROCONTROL Experimental Centre (recording site).

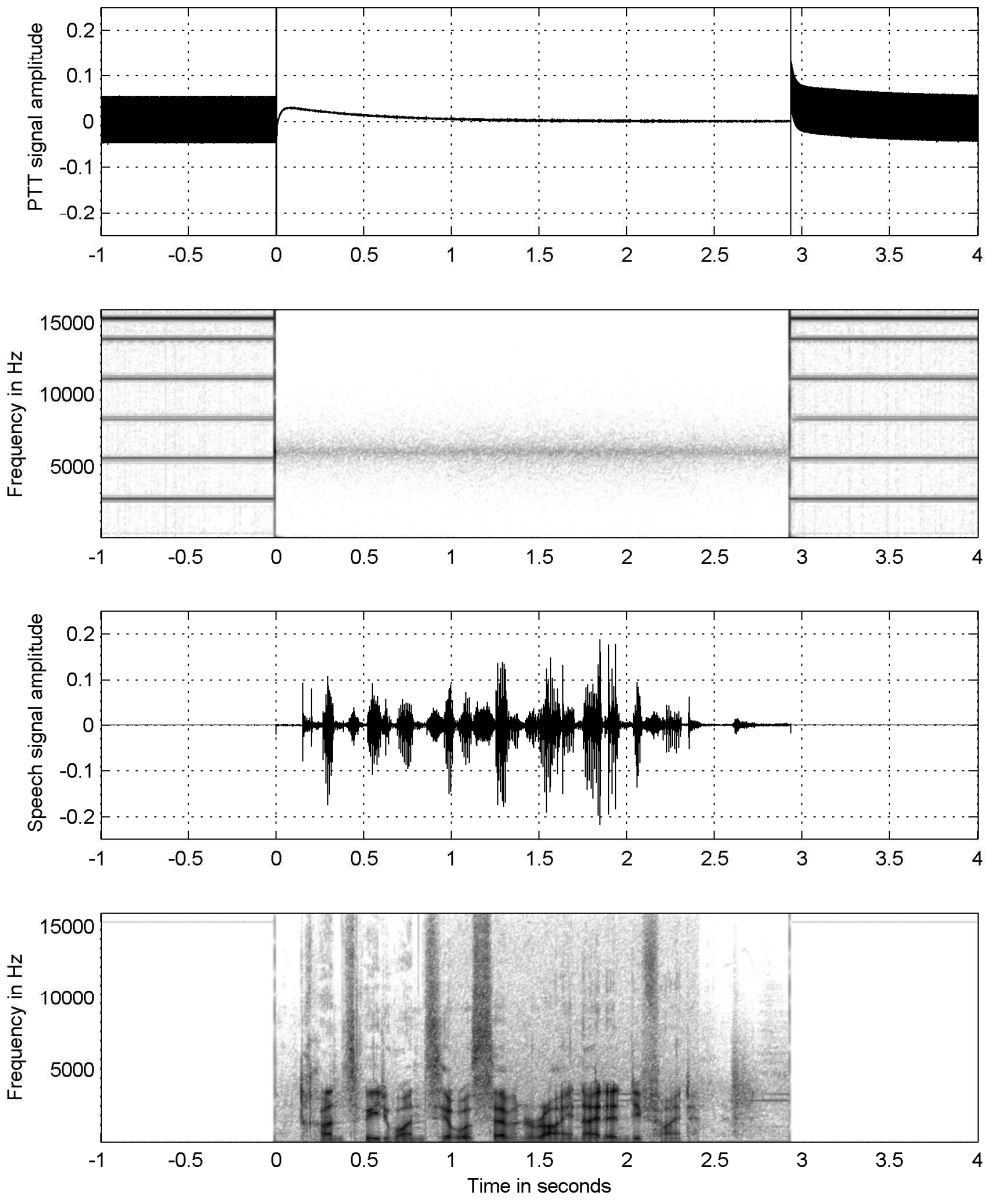


Figure 7.2.: A short speech segment (`transwede one zero seven rhein identified`) with push-to-talk (PTT) signal. Time domain signal and spectrogram of the PTT signal (top two) and time-domain signal and spectrogram of the speech signal (bottom two).

7.3. Orthographic Transcription

The speech corpus includes an orthographic transcription of the controller utterances. The orthographic transcriptions are aligned with each utterance.

7.3.1. Transcription Environment

The open-source tool TableTrans was chosen for the transcription of the corpus [127]. TableTrans was selected for its table-based input structure as well as for its capability to readily import the automatic segmentation. The transcriptionist fills out a table in the upper half of the window where each row represents one utterance. In the lower half of the window the waveform of the utterance that is currently selected or edited in the table is automatically displayed (Figure 7.3). The transcriptionist can play, pause and replay the currently active utterance by a single key stroke or as well select and play a certain segment in the waveform display. A small number of minor modifications to the TableTrans applications were made in order to lock certain user interface elements and to extend its replay capabilities.

A number of keyboard shortcuts were provided to the transcriptionist using the open-source tool AutoHotKey [128]. These were used for conveniently accessing alternative time-stretched sound files[2] and for entering frequent character or word sequences such as predefined keywords, ICAO alphabet spellings and frequent commands, both for convenience and in order to avoid typing mistakes.

7.3.2. Transcription Format

The orthographic transcription follows a strict set of rules which is given in Appendix A. In general, all utterances are transcribed word-for-word in standard British English. All standard text is written in lower-case. Punctuation marks including periods, commas and hyphens are omitted. Apostrophes are used only for possessives (e.g. `pilot's radio`)[3] and for standard English contractions (e.g. `it's`, `don't`). Numbers, letters, navigational aids and radio call signs are transcribed following a given definition based on several aeronautical standards and references. Regular letters and words are preceded or followed by special characters to mark truncations (`=`), individually pronounced letters (`~`) or unconfirmed airline names (`@`).

Stand-alone technical mark-up and meta tags are written in upper case letters with enclosing squared brackets. They denote human noises such as coughing, laughing and sighs (`[HNOISE]`), fragments of words (`[FRAGMENT]`), empty utterances (`[EMPTY]`), non-sensical words (`[NONSENSE]`), and unknown words (`[UNKNOWN]`). Groups of words are embraced by opening and closing XML-style tags to mark off-talk (`<OT> ... </OT>`), which is also transcribed, and foreign language (`<FL> </FL>`), for which a transcription could be added at a later stage. Table 7.2 gives several examples of transcribed

[2] The duration of each utterance was stretched by a factor of 1.7 using the PRAAT implementation of the PSOLA method [77]. These time-stretched sound files were used only when dealing with utterances that were difficult to understand.

[3] Corpus transcription excerpts are written in a `mono-spaced typewriter` font.

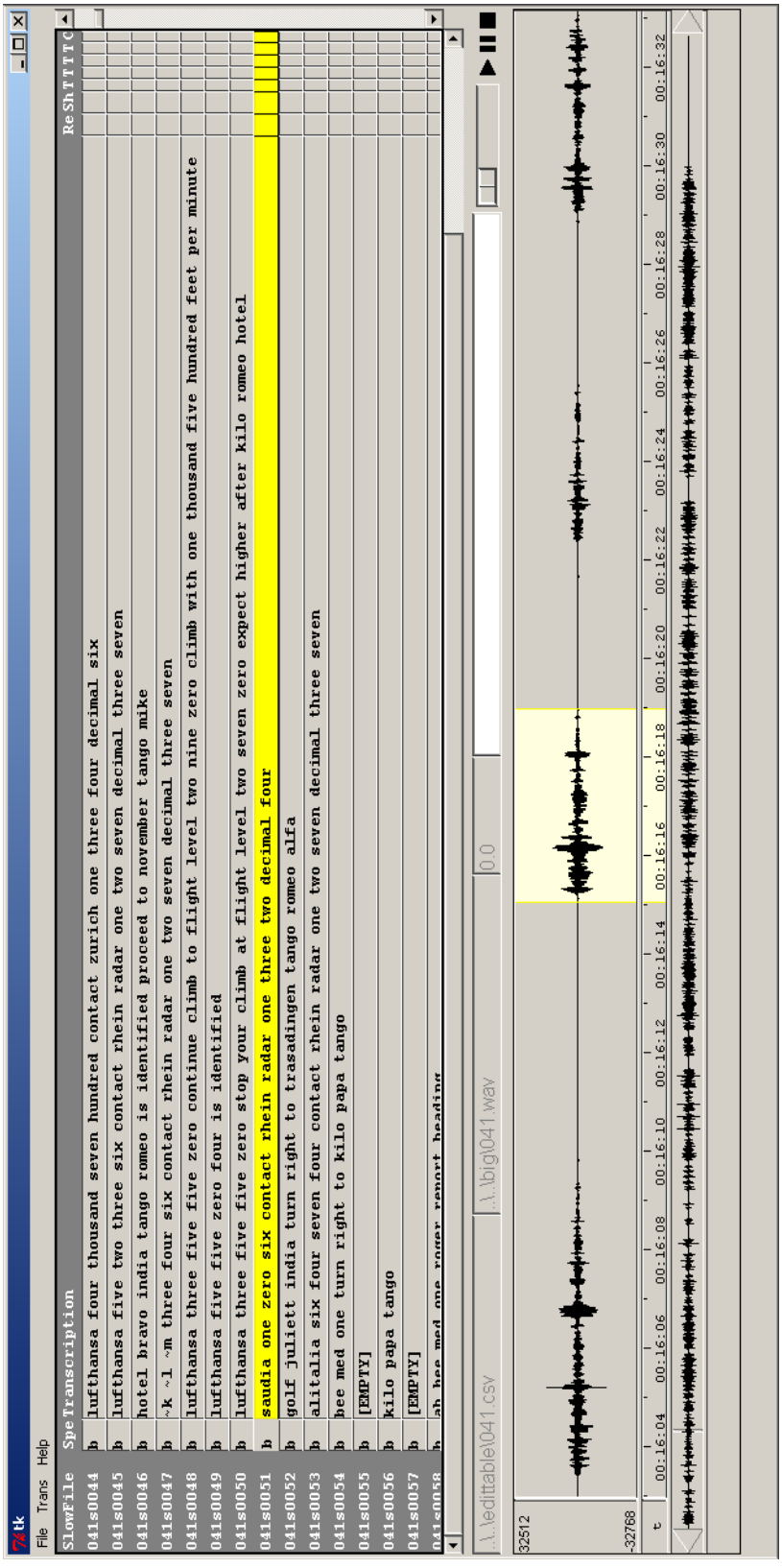


Figure 7.3.: Screen-shot of the transcription tool TableTrans.

utterances.

Silent pauses both between and within words are not transcribed. For consistent results this would require an objective measure and criterion and is thus easier to integrate in combination with a potential future word segmentation of the corpus. Also technical noises as well as speech and noises in the background—produced by speakers other than the one recorded—are not transcribed, as they are virtually always present and are part of the typical acoustical situation in an air traffic control room.

Table 7.2.: Examples of Controller Utterance Transcriptions in the ATCOSIM Corpus
Human noises are labeled with [HNOISE], word fragments with [FRAGMENT] and unintelligible words with [UNKNOWN]. Truncations are marked with an equals sign (=), individually pronounced letters with a tilde (~), and beginning and end of off-talk is labeled with <OT> and </OT>.

aero lloyd five one seven proceed direct to frankfurt
speedway three three five two contact milan one three four five two bye
ah lufthansa five five zero four turn right ten degrees report new heading
hapag lloyd six five three in radar contact climb to level two nine zero
scandinavian six one seven proceed to fribourg fox romeo india
ind= israeli air force six eight six resume on navigation to tango
good afternoon belgian airforce forty four non ~r ~v ~s ~m identified
[HNOISE] alitalia six four seven four report your heading
[FRAGMENT] hapag lloyd one one two identified
sata nine six zero one is identified <OT> oh it's over now </OT>
aero lloyd [UNKNOWN] charlie papa alfa guten tag radar contact

7.3.3. Transcription Process and Quality Assurance

The entire corpus was transcribed by a single person, which promises high consistency of the transcription across the entire database. The native English speaker was introduced to the basic ATC phraseology [113] and given lists covering country-specific toponyms and radio call signs (e.g., [129]). Clear transcription guidelines were established and new cases that were not yet covered by the transcription format immediately discussed.

Roughly three percent of all utterances were randomly selected across all speakers and used for training of the transcriptionist. This training transcription was also used to validate the applicability of the transcription format definition and minor changes were made. The transcriptions collected during the training phase were discarded and the material re-transcribed in the course of the final transcription.

After the transcription was finished, the transcriptionist once again reviewed all utterances, verified the transcriptions and applied corrections where necessary. Remaining unclear cases were shown to an operational air traffic controller and most of

them resolved.

Due to the frequent occurrence of special location names and radio call signs an automatic spell check was not performed. Instead of this, a lexicon of all occurring words was created, which includes a count of occurrence and examples of the context in which the word occurs. Due to the limited vocabulary used in ATC, this list consists of less than one thousand entries including location names, call signs, truncated words, and special mark-up codes. Every item of the list was manually checked and thus typing errors eliminated.

7.4. ATCOSIM Structure, Validation and Distribution

The entire corpus including the recordings and all meta data has a size of approximately 2.5 gigabyte and is available in digital form on a single DVD or an electronic ISO disk image. Some statistics of the corpus are given in Table 7.3.

7.4.1. Corpus Structure and Format

The ATCOSIM corpus data is composed of four directories.

The 'WAVdata' directory contains the recorded speech signal data as single-channel Microsoft WAVE files with a sample rate of 32 kHz and a resolution of 16 bits per sample. Each file corresponds to one controller utterance. The 10,078 files are located in a sub-directory structure with a separate directory for each of the ten speakers and sub-directories thereof for each simulation session of the speaker.

The 'TXTdata' directory contains single text files with the orthographic transcription for each utterance. They are organized in the same way as the audio files. The directory also contains an alphabetically sorted lexicon and a comma-separated-value file which includes not only the transcription of all utterances but also all meta data such as speaker, utterance and session IDs and transcriptionist comments.

The files in the 'HTMLdata' directory are HTML files which present the transcriptions and the meta data in a table-like format. They enable immediate sorting, reviewing and replaying of utterances and transcriptions from within a standard HTML web browser.

Last, the 'DOC' directory contains all documentation related to the corpus.

7.4.2. Validation

The validation of the database was carried out by the Signal Processing and Speech Communication Laboratory (SPSC) of Graz University of Technology, Austria. The examiner and author of the validation report has not been involved in the production of the ATCOSIM corpus, but only carried out an informal pre-validation and the formal final validation of the corpus.

The validation procedure followed the guidelines of the Bavarian Archive for Speech Signals (BAS) [130]. It included a number of automatic tests concerning completeness, readability, and parsability of data, which were successfully performed without

Table 7.3.: Key Figures of the ATCOSIM Corpus

Duration total (thereof speech)	51.4 h (10.7 h)
Data size	2.4 GB
Speakers (thereof female/male)	10 (4/6)
Sessions total	50
Sessions per speaker	7, 9, 5, 6, 1, 2, 2, 8, 7, 3
Utterances total	10078
Utterances per speaker	1167, 1848, 808, 1162, 238, 384, 378, 1716, 1739, 638
Utterance duration (mean, std. deviation, min, max)	3.8 s, 1.6 s, 0.04 s, 38.9 s
Utterances, containing	
- <FL> </FL>	182
- <OT> </OT>	84
- [EMPTY]	319
- [FRAGMENT]	35
- [HNOISE]	62
- [NONSENSE]	11
- [UNKNOWN]	11
Words	108883
Characters (thereof without space)	626425 (517542)
Lexicon entries	
- Total	858
- Meta tags	9
- Truncations	106
- Compounds	13
- Unique words	730

revealing errors. Furthermore, manual inspections of documentation, meta data, transcriptions, and the lexicon were done, which showed minor shortcomings that were fixed before the public release of the corpus. Finally, a re-transcription of 1% of the corpus data was made by the examiner, showing a transcription accuracy on word level of 99.4%, proving the transcriptions to be accurate. The ATCOSIM corpus was therefore declared to be in a usable state for speech technology applications.

7.4.3. License and Distribution

The ATCOSIM corpus is available online at `http://www.spsc.tugraz.at/ATCOSIM` and provided free of charge. It can be freely used for research and development, also in a commercial environment. The corpus is also foreseen to be distributed on DVD through the European Language Resources Association (ELRA).

7.5. Conclusions

The ATCOSIM corpus is a valuable contribution to application-specific language resources. To our best knowledge currently no other speech corpus is publicly available that contains non-prompted air traffic control speech with direct microphone recordings, as it is difficult to produce such recordings for public distribution. The large-scale real-time ATC simulations exploited for this corpus production provide an opportunity to record ATC speech which is very similar to operational speech, while avoiding the legal hassle of recording operational ATC speech.

Applications

The application possibilities for spoken language technologies in air traffic control are manifold, and the corpus can be utilized within different fields of speech-related research and development.

ATC Language Study Analysis can be undertaken on the language used by controllers and on the instructions given.

Listening Tests The speech recordings can be used as a basis for ATC-related listening tests and provide real-world ATC language samples.

Speaker Identification The corpus can be used as testing material for speaker identification and speaker segmentation applications in the context of ATC.

ATC Speech Transmission System Design The corpus can be used for the design, development and testing of ATC speech transmission and enhancement systems. Examples where the corpus was applied include the work presented in this thesis and the pre-processing of ATC radio speech to improve intelligibility [131].

Speech Recognition The corpus can also be used for the development and training of speech recognition systems. Such systems might become useful in future

ATC environments and for example be used to automatically input into the ATC system the controller instructions given to pilots .

We expand on the speech recognition example:

The controller sees among other information the current flight level of the aircraft on the radar screen in the text label corresponding to the aircraft. For example, a controller issues an instruction to an aircraft to climb from its current flight level 300 to flight level 340 (e.g. ‘‘`sabena nine seven zero climb to flight level three four zero`’’). In certain ATC display systems, the controller now enters this information (‘climb to 340’) into the system and it shows up in the aircraft label, as this information is relevant later on when routing other adjacent aircraft. However, the voice radio message sent to the pilot already contains all information required by the system, namely the aircraft call-sign and the instruction.

Depending on the achievable robustness, an automatic speech recognition (ASR) system that recognizes the controller's voice radio messages could perform various tasks: In case of extremely high accuracy the system could gather the information directly from the voice message without any user interaction. The ASR system could otherwise provide a small list of suggestions, to ease the process of entering the instructions into the system. Alternatively, the system could compare in the background the voice messages sent to the pilot and the instructions entered into the system, and give a warning in case of apparent discrepancies.

Compared to other ASR applications, the conditions would be comparably favorable in this scenario: The signal quality is high due to the use of a close-talk microphone and the absence of a transmission channel. The vocabulary is limited, and additional side information, such as the aircraft present in the sector and context-related constraints, can be exploited. The ASR system can be speaker-dependent and pre-trained, and continue training due do the constant feedback given by the controller during operation.

Possible Extensions

Depending on the application, further annotation layers might be useful and could be added to the corpus.

Phonetic Transcription A phonetic transcription is beneficial for speech recognition purposes. For accurate results it requires transcriptionists with a certain amount of experience in phonetic transcription or at least a solid background in phonetics and appropriate training.

Word and Phoneme Segmentation A more fine-grained segmentation is also beneficial for speech recognition purposes. It can often be performed semi-automatically, but still requires manual corrections and substantial effort.

Semantic Transcription A semantic transcription would describe in a formal way the actual meaning of each utterance in terms of its functionality in air traffic control, such

as clearance and type of clearance, read-back, request, ... This would support speech recognition and language interface design tasks, as well as ATC language studies. The production of such a transcription layer requires good background knowledge in air traffic control. Due to lack of contextual information, such as the pilots' utterances, certain utterances might appear ambiguous.

Call Sign Segmentation and Transcription This transcription layer marks the signal segment which includes the call sign and extracts the corresponding part of the (already existing) orthographic transcription. This can be considered as a sub-part of the semantic transcription which can be achieved with significantly less effort and requires little ATC related expertise. Nevertheless this might be beneficial for certain applications such as language interface development.

Extension of Corpus Size and Coverage With an effective size of ten hours of controller speech the corpus might be too small for certain applications. There are two reasons that would support the collection and transcription of more recordings. The first reason is the pure amount of data that is required by the training algorithms in modern speech recognition systems. The second reason is the need to extend the coverage of the corpus in terms of speakers, phonological distribution and speaking style, as well as control task and controlled area.

This chapter presented a database of ATC controller speech, which represents the input signal of the watermarking-based enhancement application considered in this thesis. Likewise, the following chapter presents a database of measurements of the ATC voice radio channel, which constitutes the transmission channel of the application at hand.

Chapter 8

Aeronautical Voice Radio Channel Measurement System and Database

This chapter presents a system for measuring time-variant channel impulse responses and a database of such measurements for the aeronautical voice radio channel. Maximum length sequences (MLS) are transmitted over the voice channel with a standard aeronautical radio and the received signals are recorded. For the purpose of synchronization, the transmitted and received signals are recorded in parallel to GPS-based timing signals. The flight path of the aircraft is accurately tracked. A collection of recordings of MLS transmissions is generated during flights with a general aviation aircraft. The measurements cover a wide range of typical flight situations as well as static back-to-back calibrations. The resulting database is available under a public license free of charge.

Parts of this chapter have been published in K. Hofbauer, H. Hering, and G. Kubin, "A measurement system and the TUG-EEC-Channels database for the aeronautical voice radio," in *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Montreal, Canada, Sep. 2006, pp. 1–5.

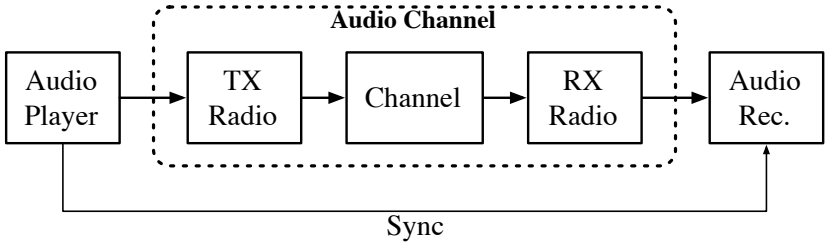


Figure 8.1.: A basic audio channel measurement system.

8.1. Introduction

A large number of sophisticated radio channel models exist, based on which the behavior of a channel can be derived and simulated [132, 133, 6]. In Appendix B, the basic concepts in the modeling and simulation of the mobile radio channel are reviewed. The propagation channel is time-variant and dominated by multipath propagation, Doppler effect, path loss and additive noise. Stochastic reference models in the equivalent complex baseband facilitate a compact mathematical description of the channel's input-output relationship. Three different small-scale area simulations of the aeronautical voice radio channel are presented in Appendix B, and we demonstrate the practical implementation based on a scenario in air/ground communication.

Unfortunately, very few measurements of the aeronautical voice radio channel that could support and quantify the theoretical models are available to the public [134]. The aim of the experiments presented in this chapter is to obtain knowledge about the characteristics of the channel through measurements in realistic conditions. These measurements shall support the insight obtained through the existing theoretical channel models discussed in Appendix B. Therefore, we measure the time-variant impulse response of the aeronautical VHF voice radio channel between the aircraft and the ground-based transceiver station under various conditions.

The outline of this chapter is as follows: Section 8.2 shows the design and implementation of a measurement system for the aeronautical voice channel. Section 8.3 gives an overview of the conducted measurement flights. The collected data form the freely available *TUG-EEC-Channels* database, which is presented in Section 8.4.

8.2. Measurement System for the Aeronautical Voice Radio Channel

Conventional wideband channel sounding is virtually unfeasible in the aeronautical VHF band. It requires a large number of frequency channels, which are reserved for operational use and are therefore not available. Thus a narrowband method is presented, which moreover allows a simpler and less expensive setup.

Figure 8.1 shows the basic concept of the proposed system: A known audio signal is transmitted over the voice radio channel and the received signal is recorded. The

time-variant impulse response of the channel can then be estimated from the known input/output pairs.

Measuring the audio channel from baseband to baseband has certain advantages and disadvantages. On the one hand, besides its simplicity, the measurement already takes place in the same domain as where our target application, a speech watermarking system, would be implemented; on the other hand, a baseband measurement does not reveal a number of parameters that could be obtained with wideband channel sounding, such as distinct power delay and Doppler spectra. Although these parameters cannot be directly measured with the proposed system, their effect on the audio channel can nevertheless be captured.

8.2.1. Synchronization between Aircraft and Ground

As already indicated in Figure 8.1, it is necessary to synchronize the measurement equipment on the transmitter and the receiver side. This linking is necessary as the internal clock frequency of two devices is never exactly the same and the devices consequently drift apart. Moreover, the clock-frequency is time-variant, due in part to its temperature-dependency. But, in a setup where one unit is on the ground and the other one is in an aircraft, the different clocks cannot be linked and a direct synchronization is not possible. This problem is common to all channel sounding systems.

In high-grade wideband channel sounders a lot of effort is put into achieving synchronization. In former days, accurately matched and temperature-compensated oscillators were aligned before measurement. Once they were separated, the systems still stayed in sync for a certain amount of time before they started drifting apart again. In modern channel sounders atomic clocks provide so stable a frequency reference that the systems stay in sync for a very long time once aligned. Such systems are readily available on the market, but come with a considerable financial expense, and are relatively large and complex in setup.

8.2.2. Synchronization using GPS Signals

The global positioning system (GPS) provides a time radio signal which is available worldwide and accurate to several nanoseconds. It can therefore also serve as an accurate frequency and clock reference [135]. Certain off-the-shelf GPS receivers output a so-called 1PPS (one pulse per second) signal. The signal is a 20 ms pulse train with a rate of 1 Hz. The rising edge of each pulse is synchronized with the start of each GPS and UTC (coordinated universal time) second with an error of less than 1 μs. This accuracy is in general more than sufficient for serving as clock reference for a baseband measurement system.

To the authors' knowledge no battery-powered portable audio player/recorder with an external synchronization input was available on the market at the time of measurements in 2006. As a consequence, direct synchronization between the transmitter and the receiver-side is in our application again not possible, even though a suitable clock reference signal would be available.

We propose in the remainder of this section a measurement method and setup with which synchronization can nevertheless be achieved by recording the clock reference signal in parallel to the received and transmitted signals and appropriate post-processing.

8.2.3. Hardware Setup

The basic structure of the measurement setup is shown in Figure 8.2. On the aircraft, the signals are transmitted and received via the aircraft-integrated VHF transceiver. The measurement audio signal is replayed by a portable CompactFlash-based audio player/recorder [136]. The transmitted and received signals are recorded with a second unit, in parallel with the 1 PPS synchronization signal which is provided by the GPS module [137]. The aircraft position, altitude, speed, heading, etc., are accurately recorded by a hand-held GPS receiver once every two seconds [138]. The setup is portable, battery-powered, rigid, and can be easily integrated into any aircraft. The only requirement is a connection to the aircraft radio headset connectors. Figure 8.3 shows the final hardware setup on-board the aircraft. An identical setup is used on the ground.

All devices are interconnected in a star-shaped manner through a custom-built interface box, which is fitted into a metal housing unit containing some passive circuitry and standard XLR and TRS connectors (Figure 8.4). The interface box provides a push-to-talk (PTT) switch remote control together with status indication and recording, and power supply, status indication and configuration interface for the GPS receiver. The box serves as the central connection point and provides potentiometers for signal level adjustments, test points for calibration and a switch to select transmit or receive signal routing. The design considers the necessary blocking of the microphone power supply voltage, the appropriate shielding and grounding for electromagnetic compatibility (EMC) as well as noise and cross-talk suppression.

8.2.4. Measurement Signal

The measurement signal consists of a continuously repeated binary maximum length sequence (MLS) of length $L = 63$ samples or $T = 7.875$ ms at a chip rate of 8 kHz [139]. This length promises a good trade-off between signal-to-noise ratio, frequency resolution and time resolution. It results in an MLS frame repetition rate of 127 Hz and an excitation of the channel with a white pseudo-random noise signal with energy up to 4 kHz. The anticipated rate of channel variation at the given maximum aircraft speed and the bandwidth of the channel are below these values [6].

The transmitted MLS sequence is interrupted once per minute by precomputed dual-tone multi-frequency (DTMF) tone pairs, in order to ease rough synchronization between transmitted and received signals. All audio files are played and recorded at a sample rate of 48 kHz and with a resolution of 16 bit. The measurement signal is therefore upsampled to 48 kHz by insertion of zeros after each sample in the time domain and low-pass interpolation with a symmetric FIR filter (using Matlab's `interp`-function). The signal is continuously transmitted during the measurements. However,

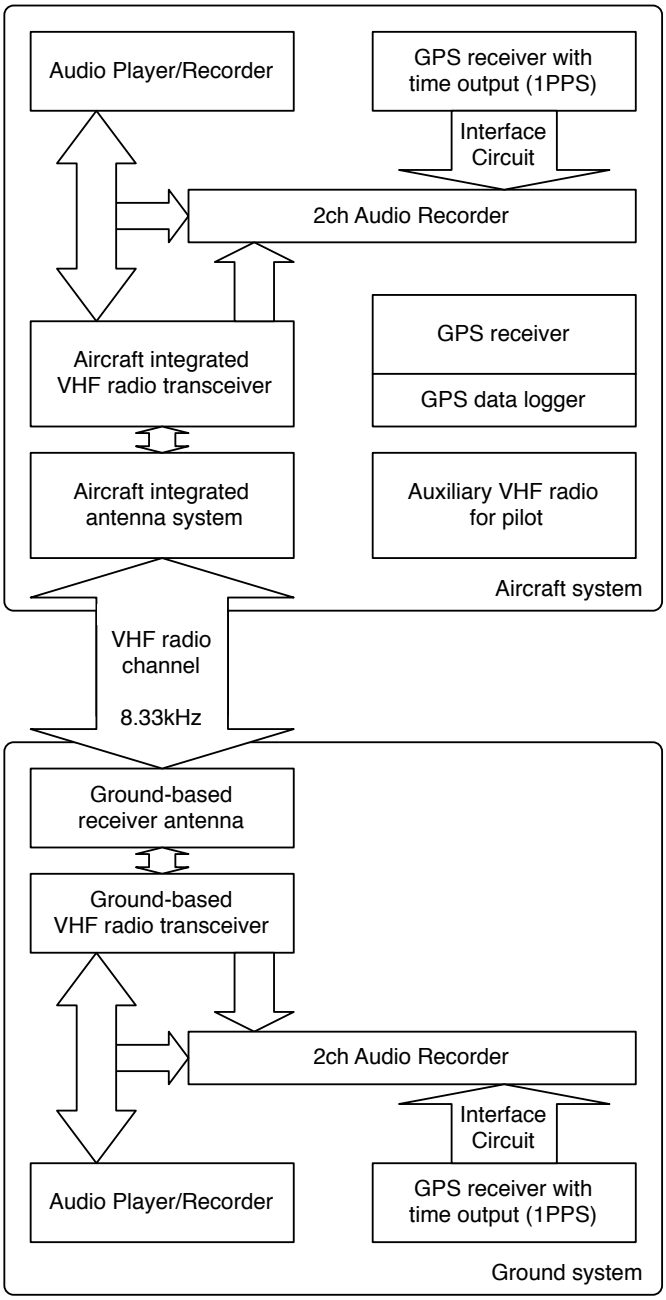


Figure 8.2.: Overview of the proposed baseband channel measurement system, using GPS timing signals for synchronization.

Figure 8.3.: Complete voice channel measurement system on-board an aircraft, with audio player and recorder, interface box, synchronization GPS (top left), and tracking GPS (bottom right).

small interruptions are made every 30 seconds and every couple of minutes due to technical and operational constraints.

8.2.5. Data Processing

The post-processing of the recorded material consists of data merging, alignment, synchronization, processing, and annotation, and is mostly automated in Matlab. The aim is to build up a database of annotated channel measurements, which is described in Section 8.4. It follows an outline of the different processing steps.

8.2.5.1. Incorporation of Side Information

The original files are manually sorted into a directory structure and labeled with an approximate time-stamp. The handwritten notes about scenarios, the transmission directions, the corresponding files, the time-stamps and durations are transferred into a data structure to facilitate automatic processing. Based on the data structure, the start time of each file is estimated.

The GPS data is provided in the widely used GPX-format and is imported into Matlab by parsing the corresponding XML structure, from which longitude, latitude, altitude and time-stamp can be extracted [140]. The remaining parameters are computed out of these values, also considering the GPS coordinates of the airport tower. The speed values are smoothed over time using robust local regression smoothing with the RLOESS method with a span of ten values or 20 s [141].

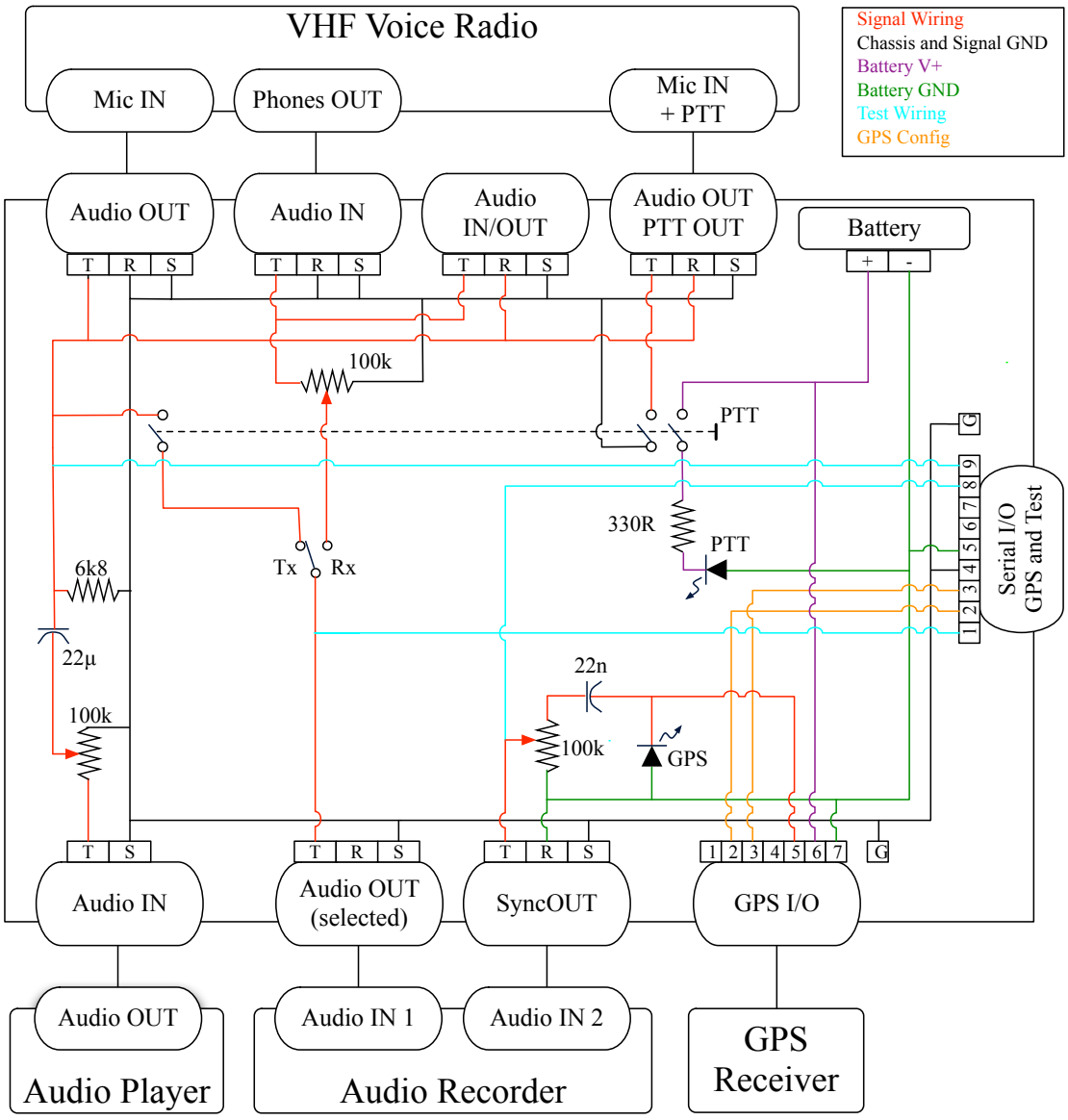


Figure 8.4.: Overview and circuit diagram of measurement system and interface box.

8.2.5.2. Analysis of 1PPS Signal and Sample Rate

The recorded files are converted into Matlab-readable one-channel Wave files using a PRAAT script [77]. From the 1PPS recording, the positions of the pulses are identified by picking the local maximum values of a correlation with a prototype shape. As it is known that the pulses are roughly one second apart, false hits can be removed, and missing pulses can be restored by interpolation. The time difference in samples between two adjacent pulses defines the effective sample rate at which the file was recorded. It was found that the sample rate is fairly constant but differs among the different devices by several Hertz. It is therefore necessary to compensate for the clock rate differences. For the aircraft and ground recordings, the sample rate can be recovered by means of the 1PPS track. Based on those the sample rate of the audio player can also be estimated.

8.2.5.3. Detection of DTMF Tones and Offset

For detecting the DTMF tones, the non-uniform discrete Fourier transform (NDFT) is computed at the specific frequencies of the DTMF tones [142]. The tones are detected in the power spectrum of the NDFT by correlation with a prototype shape, as the duration and relative position of the tones is known from the input signal. Again after filtering for false hits, the actual DTMF symbol is determined by the maxima of the NDFT output values at the corresponding position. The detected DTMF sequences of corresponding air/ground recording pairs are then aligned with approximate string matching using the Edit distance measure [108]. Their average offset in samples is computed and verified with the offset that is obtained from the data structure with the manual annotations. These rough offsets are refined using the more accurate 1PPS positions of both recordings.

8.2.5.4. Frame Segmentation and Channel Response

The recording of the received signal with a sample rate of 48 kHz is split up into frames of 378 samples, which is the length of one MLS. The corresponding frame in the recording of the transmitted signal is found by the previously computed time-variant offsets between the 1PPS locations in the two recordings, and then with the local position in-between two 1PPS pulses. Through this alignment the synchronization between transmitted and recorded signal is reestablished. The alignment is accurate to one 48 kHz sample, which is one sixth of a sample in terms of the original MLS sequence.

Based on the realigned transmission input-output pairs, the channel frequency response for every frame is estimated based on FIR system identification using the cross power spectral density (CPSD) method with

$$H(\omega_k) = \frac{\overline{X(\omega_k)} \cdot Y(\omega_k)}{|X(\omega_k)|^2}, \tag{8.1}$$

where X, Y, and H are the discrete Fourier transforms of the channel's input signal, output signal and impulse response, respectively [143]. For numerical stability the

(a) Aircraft during measurements.

(b) Aircraft transceiver.

Figure 8.5.: General aviation aircraft and onboard transceiver used for measurements.

signals are downsampled to a sample rate of 8 kHz before system identification, as the channel was excited only up to a frequency of 4 kHz.

All frames are annotated with the flight parameters and the contextual information of the meta data structure using the previously computed time-stamps and offsets. A more accurate channel model and estimation scheme based on the measurement results is presented in Chapter 9.

8.3. Conducted Measurements

A number of ground-based and in-flight measurements were undertaken with a general aviation aircraft at and around the Punitz airfield in Austria (airport code LOGG). After initial trials at the end of March 2006, measurements were undertaken for two days in May 2006.

The aircraft used was a 4-seat Cessna F172K with a maximum speed of approximately 70 m/s and a Bendix/King KX 125 TSO communications transceiver (Figure 8.5). The initial trials were undertaken with a different aircraft.

On ground, the communications transceiver Becker AR 2808/25 was used (Figure 8.6). It is permanently installed in the airport tower in a fixed frequency configuration. For reference and comparison, a Rohde&Schwarz EM-550 monitoring receiver which provides I/Q demodulation of the received AM signal was applied. The I/Q data with a bandwidth of 12 kHz was digitally recorded onto a laptop computer, however without a 1PPS synchronization signal. Synchronization is to a certain extent nevertheless possible using the DTMF markers in the transmitted signal. For the monitoring receiver a $\frac{\lambda}{4}$-ground plane antenna with a magnetic base was mounted onto the roof of a van.

The measurements occurred on the AM voice radio channel of Punitz airport at a carrier frequency of 123.20 MHz. Measurements were taken both in uplink and downlink direction, with the tower or the aircraft radio transmitting, respectively. The I/Q monitoring receiver continuously recorded every transmission.

(a) I/Q measurement receiver.

(b) Airport tower transceiver.

Figure 8.6.: Ground-based transceivers used for measurements.

A series of measurement scenarios was set up which covered numerous situations. The ground measurements consisted of static back-to-back measurements with the aircraft engine on and off. Static measurements were undertaken with the aircraft at several positions along a straight line towards the tower, spaced out by 30 cm. Additional measurements were taken with a vehicle parked next to the aircraft. The ground based measurements were concluded with the rolling of the aircraft on the paved runway at different speeds. A variety of flight maneuvers was undertaken. These covered the following situations and parameter ranges:

- Low and high altitudes (0m to 1200m above ground)
- Low and high speeds (up to 250 km/h)
- Ascends and descends
- Headings perpendicular and parallel to line-of-sight
- Overflights and circular flights around airport
- Take-offs, landings, approaches and overshoots
- Flights across, along and behind a mountain
- Flights in areas with poor radio reception (no line-of-sight)

Thorough organizational planning, coordination and cooperation between all parties involved proved to be crucial for a successful accomplishment of measurements in the aeronautical environment.

8.4. The *TUG-EEC-Channels* Database

Based on the above measurements we present an aeronautical voice radio channel database, the *TUG-EEC-Channels* database. It is available to the research community under a public license free of charge online at the following address:

`http://www.spsc.TUGraz.at/TUG-EEC-Channels/`

The *TUG-EEC-Channels* database covers the channel measurements described in Section 8.3 with three extensive flights and several ground measurements with a total duration of more than five hours or two million MLS frames. For each frame of 8 ms (corresponding to 378 samples at 48 kHz) the following data and annotation is provided:

- Time-stamp
- Transmission recording segment
- Reception recording segment
- Type of transmitted signal

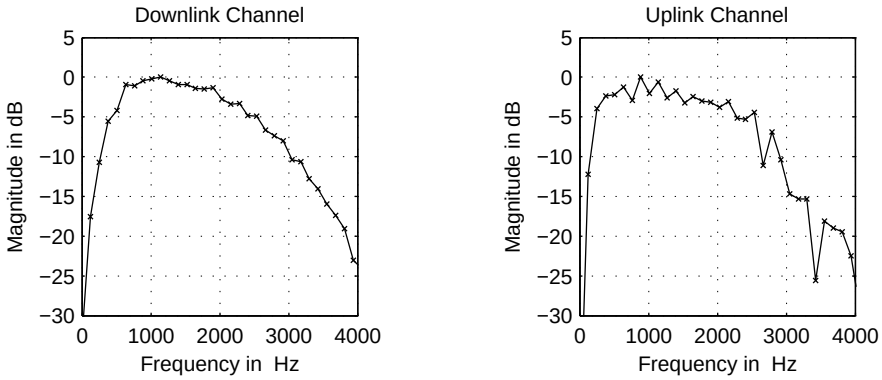


Figure 8.7.: Normalized frequency response of the static back-to-back voice radio channel in up- and downlink direction.

- I/Q recording segment
- Link to original sequence
- Uplink/downlink direction
- Estimated channel impulse response
- Flight parameters elevation, latitude, longitude, distance to tower, speed, and azimuth relative to line-of-sight
- Experimental situation (engine on/off, aircraft on taxiway or flying, ...)
- Plain-text comments

An exact description of the data format of the *TUG-EEC-Channels* and additional information can be found on the website. Upon request, the raw MLS recordings as well as a number of voice signal transmission recordings and voice and background noise recordings made in the cockpit using conventional and avionic headset microphones can also be provided.

The following figures present several excerpts of the database. Figure 8.7 shows the magnitude response of the transmission chain in both directions, which is mostly determined by the radio transceivers. In Figure 8.8, the magnitudes of selected frequency bins of the downlink channel are plotted over time. The basic shape of the frequency response is maintained, but the magnitude varies over time. This time variation results from a synchronization error which is compensated in the more accurate model of Chapter 9. Figure 8.9 illustrates a GPS track.

8.5. Conclusions

We presented a system for channel measurements and the *TUG-EEC-Channels* database of annotated channel impulse and frequency responses for the aeronautical VHF voice

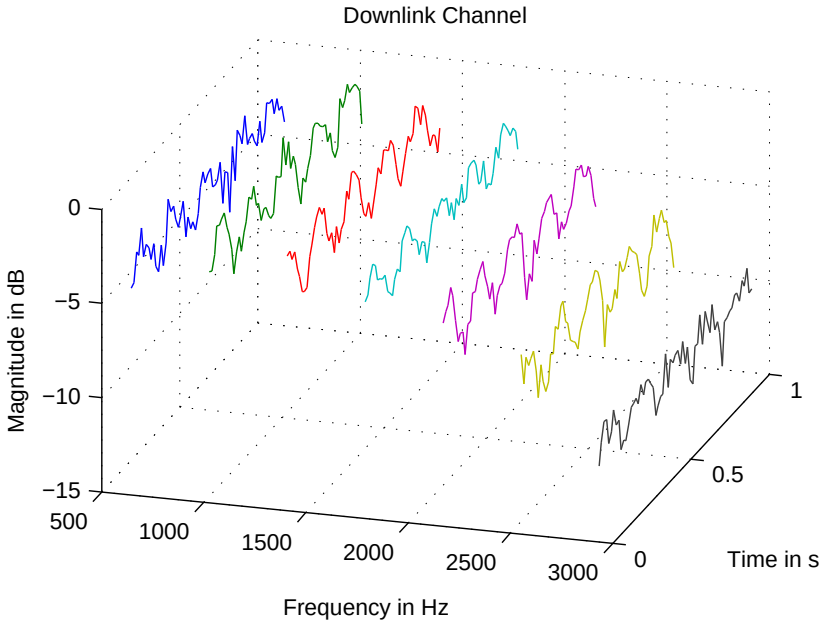


Figure 8.8.: Evolution of the frequency bins at 635, 1016, 1397, 1778, 2159, 2540 and 2921 Hz over time. Aircraft speed is 44.7 $\frac{\text{m}}{\text{s}}$ at a position far behind the mountain and an altitude of 1156 m.

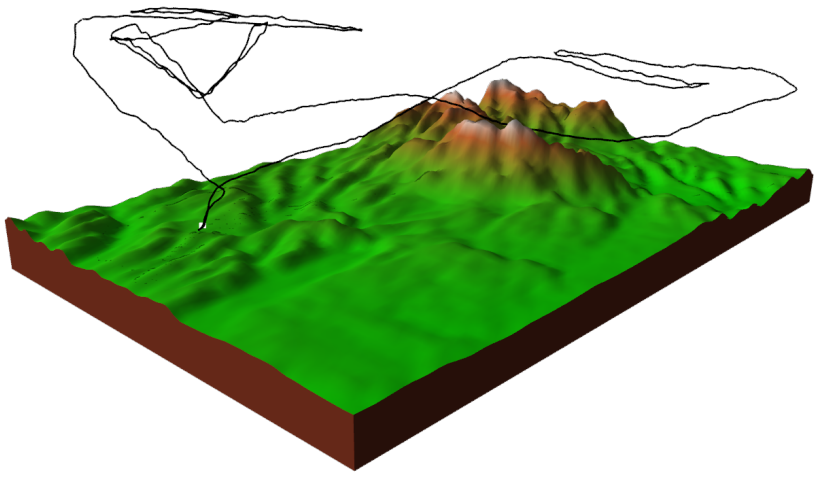


Figure 8.9.: Visualization of the aircraft track based on the recorded GPS data.

channel. The current data can be directly applied to simulate a transmission channel for system test and evaluation. We hope that the aeronautics community will make extensive use of our data for future developments, which would also help to verify the validity of our measurement scenarios.

Beyond that, it would be interesting to see what conclusions with respect to parameters of the aeronautical radio channel models can be drawn from the data. In Chapter 9, we present a further analysis and abstraction of the data and propose a model of the aeronautical VHF voice channel based on the collected data.

Acknowledgments

We are grateful to FH JOANNEUM (Graz, Austria), and Dr. Holger Flühr and Gernot Knoll in particular, for taking charge of the local organization of the measurements. We also want to thank the Union Sportfliegerclub Punitz for the kind permission to use their facilities, Flyers GmbH and Steirische Motorflugunion for providing the aircraft, and Rohde & Schwarz for contributing the monitoring receiver. Finally, we want to recognize all the other parties that helped to make the measurements possible. The campaign was funded by EUROCONTROL Experimental Centre (France).

Chapter 9

Data Model and Parameter Estimation

In this chapter, we propose a data model to describe the data in the *TUG-EEC-Channels* database, and a corresponding estimation method. The model is derived from various effects that can be observed in the database, such as different filter responses, a time-variant gain, a sampling frequency offset, a DC offset and additive noise. To estimate the model parameters, we compare six well-established FIR filter identification techniques and conclude that best results are obtained using the method of Least Squares. We also provide simple methods to estimate and compensate the sampling frequency offset and the time-variant gain.

The data model achieves a fit with the measured data down to an error of -40 dB, with the modeling error being smaller than the channel's noise. Applying the model to select parts of the database, we conclude that the measured channel is frequency-nonselective. The data contains a small amount of gain modulation (flat fading). Its source could not be conclusively established, but several factors indicate that it is *not* a result of radio channel fading. The observed noise levels are in a range from 40 dB to 23 dB in terms of SNR.

This chapter presents recent results that have not yet been published.

9.1. Introduction

The *TUG-EEC-Channels* database contains an estimated channel impulse response for every measurement frame based on the cross power spectral density of the realigned channel input and output signal of the corresponding frame. This set of impulse responses can itself form a channel description based on the assumption of a time-variant filter channel model. We will apply this channel description, among others, in the evaluation of the watermarking system in Chapter 10.

A manual inspection of the recorded signals and the estimated channel responses revealed that the simple time-variant filter channel model underlying the CPSD estimates of (8.1) may not be an adequate representation for the input/output relationship of the data and may lead to wrong conclusions. We found the following signal properties that are not represented in the basic model.

Additive Noise The recorded signals are corrupted by additive noise, which fully affects the CPSD channel estimates.

DC Offset There are significant time-variant DC offsets in the recorded signals and the estimated impulse responses.

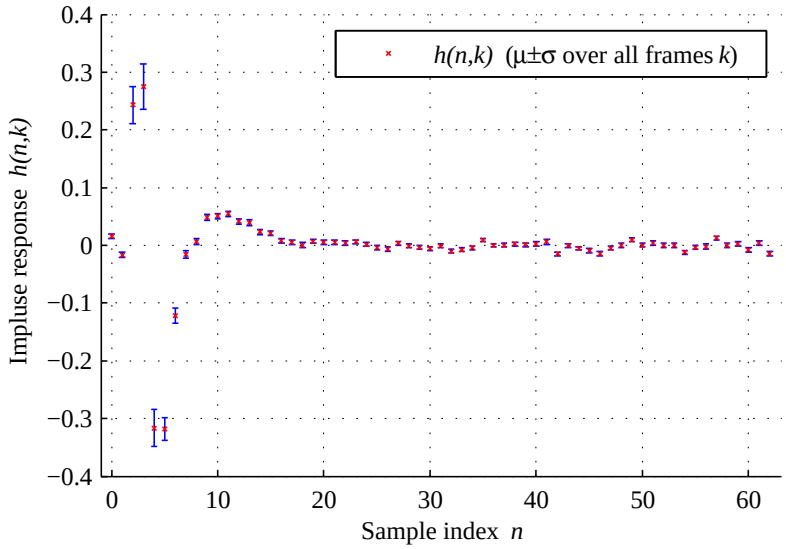
Time Variation There is a systematic time-variation of the impulse response estimates of (8.1), even in the case of the aircraft not moving. Figure 9.1 shows this for a short segment of a static back-to-back measurement with the aircraft on ground (frame number 2012056 to 2013055). We denote the n'th sample of the estimated impulse response of the k'th frame with $h(n,k)$. The DC offset of each impulse response has been removed a-priori by subtracting the corresponding mean value.

Sampling Frequency Offset Albeit small, there is a sampling frequency offset between the recordings of the transmitted and received signals.

Gain Modulation Measured in a sliding rectangular window with a length of one frame (i.e., 63 samples or 378 samples at a sample rate of 8 kHz or 48 kHz, respectively), the transmitted signal is due to its periodicity of exactly constant power. Measuring the recordings of the received signal with the same sliding window reveals an amplitude (gain) modulation, which appears to be approximately sinusoidal with a frequency of 40–90 Hz.

In the remainder of this chapter, we propose an improved channel model and a corresponding estimation method that provides stable and accurate estimations with low errors, addresses the irregularities in the channel estimates contained in the database, and provides a mean to estimate the noise in the measured channel by incorporating neighboring frames. We apply the method to select parts of the measurement database to demonstrate the validity of the model and to estimate its parameters.

(a) Mean and standard deviation of the coefficients over all frames.



(b) Time variation of selected coefficients over all frames.

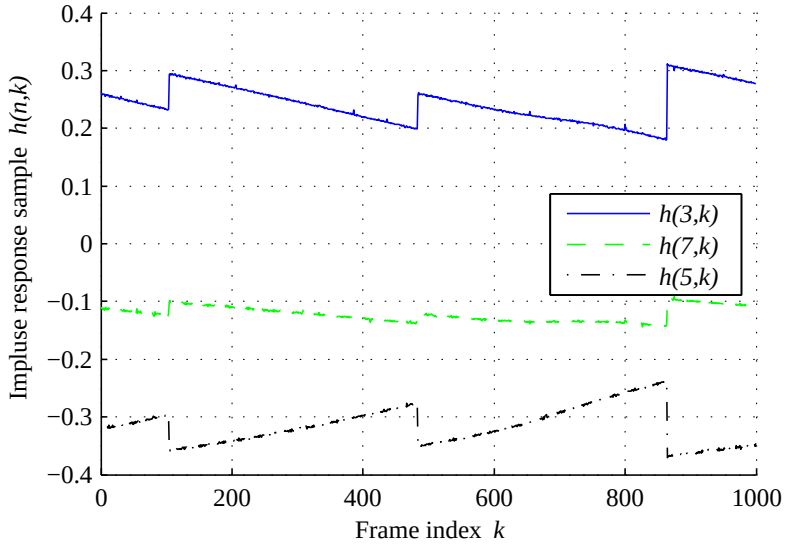


Figure 9.1.: Time-variation of the estimated impulse responses $h(n,k)$ in the TUG-EEC-Channels database for a static back-to-back measurement over 1000 frames (8 s).

Figure 9.2.: Separation between voice radio channel model and measurement errors.

9.2. Proposed Data and Channel Model

We propose an improved channel model that better represents the measured data by incorporating the aforementioned effects.

9.2.1. Assumptions

It is important to take into consideration that part of the observed effects might be measurement errors. This is represented in Figure 9.2 by separating the measurement error and voice radio channel models. It is often difficult to conclusively establish if an observed effect is caused by the measurement system or the radio channel, and we make the following assumptions.

By specification, the voice radio channel includes the radio transmitter, the radio receiver, and the VHF wave propagation channel. As such, it is a bandpass channel and does not transmit DC signal components. We attribute the DC offset to deficiencies in the measurement hardware, and to the interface box circuits in particular, and consider the time-variant DC offset (measured within a window of one measurement frame) as a measurement error.

We assume that the filtering characteristics of the system can be mostly attributed to the voice radio channel, since in laboratory measurements the frequency response of the measurement system was verified to be flat compared to the observed filtering. It appears that the filtering is dominated by the radio transmitter and receiver filters. One can observe two particular filter characteristics in the database that solely depend on the transmission direction (air/ground or ground/air) and, as such, on the transmitter and receiver used.

The sampling frequency offset can be clearly attributed to the measurement system, since the sampling clocks of the audio devices were not synchronized and the signals replayed and recorded with slightly offset sampling frequencies.

We conjecture—and later results will confirm—that the systematic time-variation of the *TUG-EEC-Channels* estimates shown in Figure 9.1 results from an insufficient synchronization between the recordings of the transmitted and the received signals. The database aims to synchronize all signal segments, no matter their content, using the GPS timing signals. It successfully does so with an accuracy of one sixth of a sample in terms of the original MLS chip rate. However, the results will show that this accuracy is not sufficient for obtaining low estimation errors. In fact, the constant slope in the lower plot of Figure 9.1 results from a slight drift between the channel input/output frame pairs. When the offset between a pair is larger than half a sample

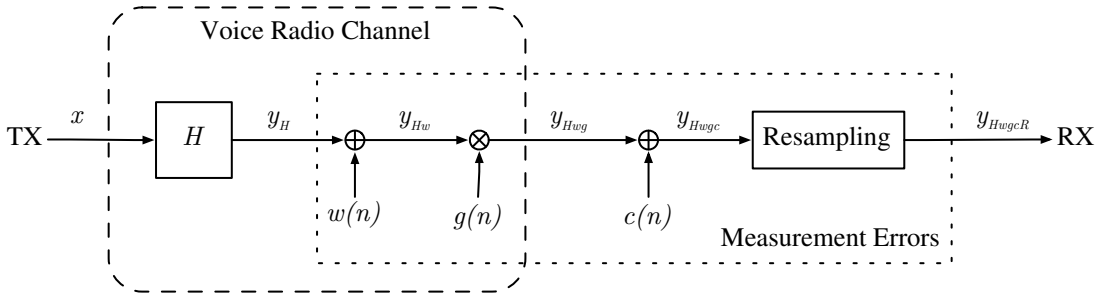


Figure 9.3.: Proposed data and channel model with a linear and block-wise time-invariant filter H, additive noise w, time-variant gain g and time-variant DC offset c. The subscripts of the signal y denote the model components that the signal has already passed. For example, y_{Hw} denotes the input signal after having passed the filter H and the addition of the noise w.

(at a sample rate of 48 kHz), the offset is automatically compensated by shifting the input by one sample, which leads to the discontinuities in $h(n,k)$ visible in Figure 9.1.

The observed gain modulation could be a measurement effect or a channel property, and we will discuss this issue in more detail at the end of this chapter.

We expect that most of the additive noise results from the voice radio channel. Nevertheless, the measurement system also contributes to the overall noise level. For example, there is small cross-talk between the recorded MLS and GPS signals due to deficiencies in the measurement hardware, and there is electromagnetic interference from the aircraft avionics.

9.2.2. Proposed Model

Based on the above assumptions, we propose a data model as shown in Figure 9.3. We merge the transmitter- and receiver-side measurement errors into one block. The overlap between the voice radio channel model and the measurement error model represents the uncertainty in the attribution of the components in the overlapping segment to one or the other category. Also, there is an uncertainty in the order of the model elements, and the depicted arrangement represents one possible choice.

9.3. Parameter Estimation Implementation

Before Section 9.3.3 shows the implementation of an estimation method for the channel model shown in Figure 9.3, we first address two sub-problems. We present a simple sampling clock synchronization method using resampling (Section 9.3.1), and compare different methods to estimate a linear and time-invariant channel filter given a known channel input and a noise-corrupted channel output (Section 9.3.2).

9.3.1. Self-Synchronization of the Received Signal by Resampling

In the following, we focus on signal segments in the database containing MLS signals, and present a simple resynchronization method in which we leave aside the transmitted signal recordings, and accurately resample the received signal recordings to the chip rate of the original MLS. While we do not claim any optimality herein, the presented method is simple and provides sufficient performance given the task at hand.

A manual inspection of the recordings showed that the sampling frequencies of the different devices are different among each other, but stable within segments of several minutes. In the following we assume that the sampling frequencies are constant within one block. A block denotes a recording fragment that contains a continuous MLS signal, and typically has a length of 20 s to 30 s.

The original MLS signal used in the measurements has a length of 63 chips at a chip-rate of 8000 chips/s. Since the signal was upsampled to a sampling frequency $f_0 = 48\,\text{kHz}$, the MLS signal is periodic with a signal period $N_0 = 6 \cdot 63 = 378$ samples or $T_0 = \frac{378}{48000}\text{s} \approx 7.875\,\text{ms}$. The signal was replayed at an unknown sampling frequency close to 48 kHz, transmitted over an analog channel, and recorded at a sampling frequency that is again close to 48 kHz but unknown.

In order to compensate for the two unknown sampling frequencies, we transform the recorded received signal $y_1(n)$, denoted RX in the database and corresponding to y_{HwgcR} in Figure 9.3, into a cubic spline representation and resample the spline with a sampling frequency f_2 such that the signal period of the resulting $y_2(n)$ is again exactly N_0 (or T_0, respectively). We determine the optimal f_2 or, equivalently, the resampling factor $\gamma = \frac{f_2}{f_0}$ using two autocorrelation-based periodicity measures and a three-step local maximization procedure.

Let the vector

$$\mathbf{x}_k = [y_2(kN_0), y_2(kN_0+1), y_2(kN_0+2), \ldots, y_2(kN_0+N_0-2), y_2(kN_0+N_0-1)]$$

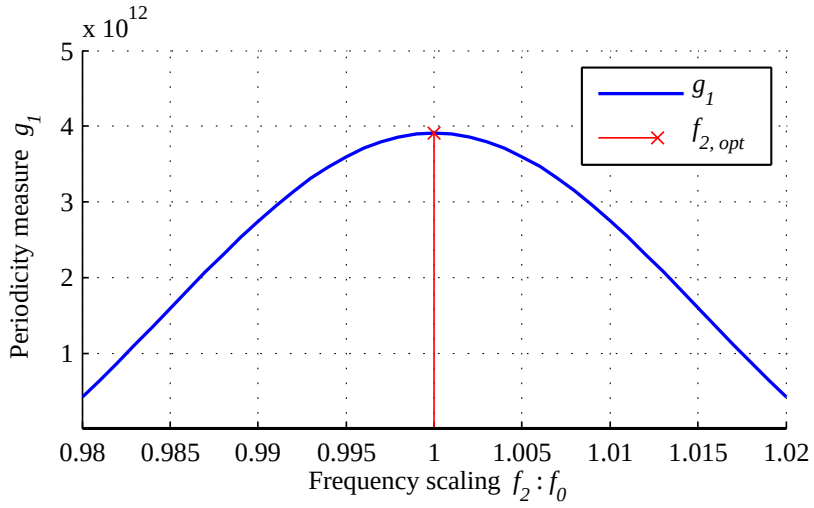
denote the k'th frame of the resampled signal (with k in the range from $0 \ldots M$-1, assuming M frames within one block of $y_1(n)$, and M even). We define two scalar error functions

$$g_1(\gamma) = \left[\begin{array}{cccccc} \mathbf{x}_0 & \mathbf{x}_2 & \ldots & \mathbf{x}_{M-4} & \mathbf{x}_{M-2} \end{array}\right] \left[\begin{array}{cccccc} \mathbf{x}_1 & \mathbf{x}_3 & \ldots & \mathbf{x}_{M-3} & \mathbf{x}_{M-1} \end{array}\right]^{\mathrm{T}}$$
$$g_2(\gamma) = \left[\begin{array}{cccccc} \mathbf{x}_0 & \mathbf{x}_1 & \ldots & \mathbf{x}_{M/2-2} & \mathbf{x}_{M/2-1} \end{array}\right] \left[\begin{array}{cccccc} \mathbf{x}_{M/2} & \mathbf{x}_{M/2+1} & \ldots & \mathbf{x}_{M-2} & \mathbf{x}_{M-1} \end{array}\right]^{\mathrm{T}}.$$

The functions g_1 and g_2 are similar to an autocorrelation of y_2 at lags N_0 and $\frac{N_0 M}{2}$, and are plotted in Figure 9.4 for an exemplary block of the database (frame-ID 2011556 to 2013057). While g_1 has a smooth and parabolic shape, which is advantageous for numeric maximization, its maximum peak is very wide and thus prone to the influence of noise. In contrast, g_2 has a very narrow and distinct peak, but exhibits many local maxima.

We maximize g_1 and g_2 as a function of the resampling factor $\gamma = \frac{f_2}{f_0}$ in the following way:

(a) Periodicity Measure g_1



(b) Periodicity Measure g_2

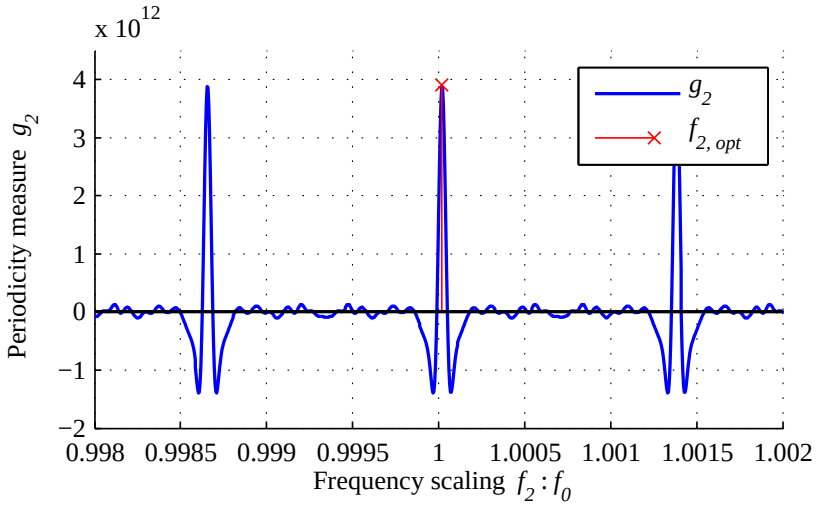


Figure 9.4.: Two periodicity measures as a function of the resampling factor $\frac{f_2}{f_0}$.

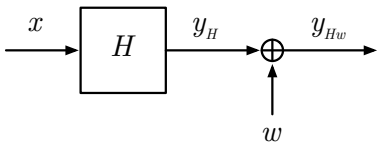


Figure 9.5.: LTI filter plus noise channel model.

1. Maximization of g_1 over γ in the interval $\gamma = [0.98; 1.02]$ using golden section search and parabolic interpolation (Matlab function 'fminbnd', [144]), resulting in the estimate γ_1.

2. Maximization of g_2 over γ in the interval $\gamma = [\gamma_1 - 2\delta; \gamma_2 + 2\delta]$ using a manually tuned $\delta = 5 \cdot 10^{-5} + \frac{0.06}{M}$ and a grid search with a search interval of 0.1δ, resulting in the estimate γ_2.

3. Maximization of g_2 over γ in the interval $\gamma = [\gamma_2 - \delta; \gamma_2 + \delta]$, again using golden section search and parabolic interpolation, and resulting in the final estimates γ_{opt} and $f_{2,\text{opt}}$ at which the spline is resampled to form $y_2(n)$.

The frequency offset $f_2 - f_0$ for different measurement signal segments is included in Table 9.2.

9.3.2. Comparison of Filter Estimation Methods

We compare six well-established methods to estimate the channel impulse response of the filter component H of the model in Figure 9.3, in terms of accuracy and noise robustness.[1]

Filter Plus Noise Channel Model

For the comparison of the different estimators, we assume a linear and time-invariant (LTI) filter plus noise model to characterize the relation between input and output. The model is characterized by the filter's impulse response $h(n)$ or frequency response $H(e^{j\omega})$, the input signal $x(n)$, the output signal $y_{Hw}(n)$ and additive white Gaussian noise (AWGN) $w(n)$ (see Figure 9.5). We denote with P the order of the filter, with N the number of input and output samples used for the channel estimation, and with the circumflex or hat accent ($\hat{}$) estimated variables.

Estimation Methods

We compare the estimation accuracy of the following methods.

[1] Parts of this subsection are based on M. Gruber and K. Hofbauer, "A comparison of estimation methods for the VHF voice radio channel," in *Proceedings of the CEAS European Air and Space Conference (Deutscher Luft- und Raumfahrtkongress)*, Berlin, Germany, Sep. 2007. The experiments were performed by Mario Gruber in the course of a Master's thesis [145].

Deconvolution The filter's impulse response $\hat{h}(n)$ is estimated recursively using [146]

$$\hat{h}(p) = \frac{1}{x(0)} \left(y_{Hw}(p) - \sum_{m=1}^{p} x(m)\hat{h}(p-m) \right).$$

Spectral Division The impulse response $\hat{h}(n)$ is estimated by an inverse discrete Fourier transformation (DFT) of the frequency response $\hat{H}(k)$, which is obtained using [146]

$$\begin{array}{cc} \text{Time domain} & \text{Frequency domain} \\ y_{Hw}(n) = x(n) * \hat{h}(n) & \circ\!\!-\!\!\bullet \quad Y(k) = X(k)\hat{H}(k) \end{array}.$$

Power Spectral Densities The frequency response $\hat{H}(k)$ can also be obtained using [146]

$$\begin{array}{cc} \text{Time domain} & \text{Frequency domain} \\ R_{xy}(n) = R_{xx}(n) * \hat{h}(n) & \circ\!\!-\!\!\bullet \quad G_{xy}(k) = G_{xx}(k)\hat{H}(k), \end{array}$$

where $R_{xx}(n) = \sum_{m=0}^{N-1} x(m)x(m+n)$ is the auto-correlation of the input $x(n)$ and $R_{xy}(n) = \sum_{m=0}^{N-1} x(m)y_{Hw}(m+n)$ the cross-correlation between input $x(n)$ and the output $y_{Hw}(n)$, and $G_{xx}(k)$ and $G_{xy}(k)$ the corresponding power spectral density (PSD) and cross-PSD.

Method of Least Squares The impulse response vector $\hat{\mathbf{h}}$ is given by the solution of the normal equation [147]

$$\hat{\mathbf{h}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}. \tag{9.1}$$

Using the covariance windowing method, the Toeplitz structured convolution matrix $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} x(P-1) & \cdots & x(1) & x(0) \\ x(P) & \cdots & x(2) & x(1) \\ \vdots & \ddots & \vdots & \vdots \\ x(N-1) & \cdots & x(N-P+1) & x(N-P) \end{bmatrix}$$

and the output vector $\mathbf{y}$ is

$$\mathbf{y} = \begin{bmatrix} y_{Hw}(\frac{P-1}{2}) \\ y_{Hw}(1+\frac{P-1}{2}) \\ \vdots \\ y_{Hw}(N-1-\frac{P-1}{2}) \end{bmatrix}.$$

Maximum Length Sequence If the input signal $x(n)$ is a maximum length sequence (MLS), then $R_{xx}(n) \approx \delta(n)$ and the cross-correlation $R_{xy}(n)$ between input $x(n)$ and output $y_{Hw}(n)$ approximates the channel's impulse response, or [148]

$$\hat{h}(n) = \delta(n) * \hat{h}(n) \approx R_{xx}(n) * \hat{h}(n) = R_{xy}(n).$$

The approximation $R_{xx}(n) \approx \delta(n)$ is only valid for long MLS. The approximation error can be fully compensated with [148]

$$\hat{h}(n) = \frac{1}{L+1}\left(R_{xy}(n) + \sum_{m=0}^{P-1} \hat{h}(m)\right).$$

Estimation Performance Comparison

Experimental Settings We estimate $P = 63$ filter coefficients using as input $N = 2^{14}$ samples of an audio signal with a sample rate $f_s = 22050\,\text{Hz}$ or a repeated MLS signal. The signal $x(n)$ is filtered using an FIR lowpass filter of order 60 ($P_{\text{LP}} = 61$) and a cutoff-frequency of 4.4 kHz, and AWGN $w(n)$ added at an SNR of 10 dB, with

$$\text{SNR} = 20\log_{10}\left(\frac{(x(n) * h(n))_{\text{RMS}}}{w_{\text{RMS}}}\right)\ \text{dB}.$$

The index RMS indicates the root mean square value of the corresponding signal.

Experimental Results We use as error measure the error in the filtered signal obtained with the original and the estimated filter. It is given by

$$e_y(n) = y_H(n) - \hat{y}_H(n) = x(n) * h(n) - x(n) * \hat{h}(n)$$

and expressed as the ratio

$$E_y = 20\log_{10}\left(\frac{e_{y,\,\text{RMS}}}{y_{H,\text{RMS}}}\right)\ \text{dB}.$$

The estimation error for all methods and for the two input signals with and without noise $w(n)$ is shown in Table 9.1. In the presence of noise the method of Least Squares outperforms all other methods. This result is consistent with estimation theory, as the method of Least Squares is an efficient estimator (i.e., it attains the Cramer-Rao lower bound) given the linear model and additive white Gaussian observation noise [147].

Estimation Parameters and Noise Estimation The number of estimated filter coefficients, so far set to $P = 63$, influences the observed estimation error. Figure 9.6 shows that in the noise-free case the estimation is error-free as soon as $P > P_{\text{LP}}$ (using the same experimental settings as above, and Least Squares estimation). In the presence of noise, the estimation error has a minimum at $P \approx P_{\text{LP}}$, since for large P the estimator begins to suffer more from the channel noise due to the random fluctuations in the redundant filter coefficient estimates.

The number of input samples, so far set to $N = 2^{14}$ samples, affects the observed estimation error. Assuming a time-invariant channel and using the same experimental settings as before, the estimation accuracy increases with increasing N (Figure 9.7). In the case of large N (i.e., $N \gg P$) and using the method of Least Squares, it is possible

Table 9.1.: Filter Estimation Results

Estimation error to output ratio E_y for different estimation methods, input signals, and with and without observation noise w.

Estimation Method	Audio	Audio+Noise	MLS	MLS+Noise
Deconvolution	< -150 dB	> +100 dB	< -150 dB	> +100 dB
Spectral Division	< -150 dB	-16 dB	-68 dB	> +100 dB
PSD	< -150 dB	-21 dB	< -150 dB	+8 dB
Least Squares	< -150 dB	-34 dB	< -150 dB	-34 dB
MLS	n/a	n/a	-27 dB	-9.7 dB
MLS (compensated)	n/a	n/a	< -150 dB	-10 dB

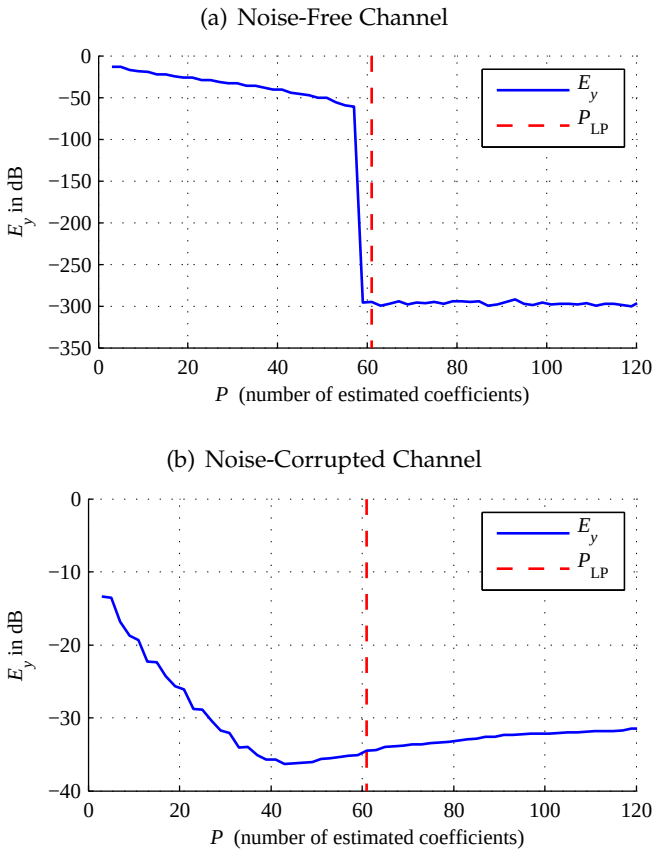


Figure 9.6.: Estimation error to output ratio E_y as a function of the number of estimated coefficients. The original filter has $P_{LP} = 61$ coefficients.

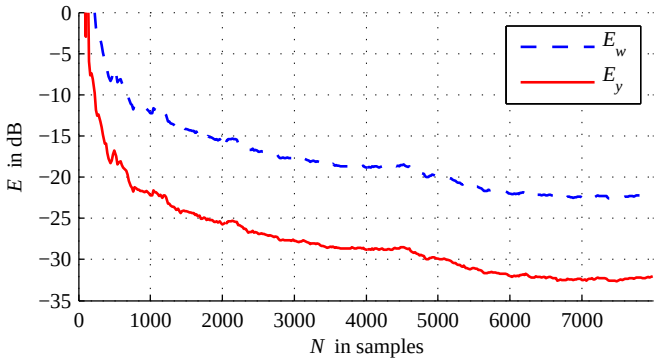


Figure 9.7.: Signal and noise estimation errors E_y and E_w at an SNR of 10 dB as a function of the number N of input and output samples ($f_s = 22050$ Hz) used for estimation.

to estimate not only the filter coefficients $h(n)$, but also the channel noise $w(n)$. The noise estimate

$$\hat{w}(n) = y_{Hw}(n) - \hat{y}_H(n) = y_{Hw}(n) - x(n) * \hat{h}(n) \tag{9.2}$$

has an estimation error

$$e_w(n) = w(n) - \hat{w}(n),$$

which is again expressed as an estimation error ratio

$$E_w = 20 \log_{10} \left(\frac{e_{w,\,\mathrm{RMS}}}{w_{\mathrm{RMS}}} \right) \ \mathrm{dB}$$

and shown in Figure 9.7 as a function of the number of samples used for estimation.

We conclude that, in comparison to the frame-based PSD estimator used for the database, the Least Squares method is a more suitable estimator. In combination with using long signal segments it also allows to estimate the channel noise, which is needed in the channel estimation method presented hereafter.

9.3.3. Channel Analysis using Resampling, Pre-Filtering, Gain Normalization and Least Squares Estimation

Based on the results of the previous two subsections, we have developed an estimation method for the channel model of Figure 9.3. We focus only on measurement data blocks containing MLS signals.

As depicted in Figure 9.8, the estimation consists of the following steps.

1. The recorded received signal $y_{HwgcR}(n)$, denoted RX in the database, is resampled with the resampling factor γ_{opt} to 48000 Hz, using the method described in Section 9.3.1.

2. The resampled signal is decimated (including anti-alias filtering) by a factor of six to the chip-rate of the original MLS sequence.

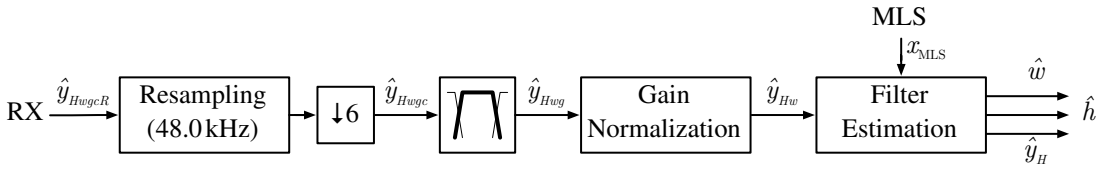


Figure 9.8.: Estimation of the channel's impulse response, time-variant gain, and additive noise based on measured data and with compensation for measurement errors.

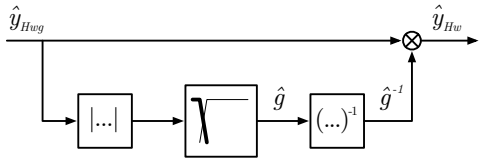


Figure 9.9.: Implementation of the gain normalization.

3. The DC offset $c(n)$ is estimated using a lowpass filter of order 1000 and a cutoff frequency of 40 Hz.

4. To remove the DC offset, and low and high frequency noise outside the frequency range of interest, $\hat{y}_{Hwgc}(n)$ is bandpass filtered using a bandpass filter with a passband from 100 Hz to 3800 Hz.

5. The time-variant gain $g(n)$ is estimated with a conventional envelope follower (shown in Figure 9.9, and using a lowpass filter with a cutoff frequency of 100 Hz). The signal $\hat{y}_{Hwg}(n)$ is then normalized to unit gain using the gain estimate $\hat{g}$.

6. The channel filter's impulse response $h(n)$ is estimated using the original binary-valued MLS signal $x_{\mathrm{MLS}}(n)$ of the database as input, $\hat{y}_{Hw}(n)$ as noise-corrupted output, and the method of Least Squares for estimation, as discussed in Section 9.3.2 and computed with Matlab's backslash operator.[2]

7. The channel noise $w(n)$ is estimated using (9.2) and

$$\hat{w}(n) = \hat{y}_{Hw}(n) - x_{\mathrm{MLS}}(n) * \hat{h}(n) = \hat{y}_{Hw}(n) - \hat{y}_H(n).$$

9.4. Experimental Results and Discussion

To show the validity of the analysis system and to obtain insight in the channel properties, we apply the analysis system presented in the previous section to select parts of the database that cover a wide range of experimental conditions.

[2]It is assumed that the filter is time-invariant within one block. Section 9.4 will show that the assumption of a time-variant filter in combination with a tracking estimation method leads to overall inferior results.

9.4.1. Filter and Noise Estimation

The noise estimate $\hat{w}(n)$, which is at the same time the error signal of the Least Squares estimation, is a strong measure for the suitability of the entire channel estimation scheme. Assuming $N \gg P$, the error signal $\hat{w}(n)$ is expected to contain no MLS signal components, but only noise. Thus, we use as error measure the overall power of the error signal $\hat{w}(n)$, as well as its spectral and acoustical quality. If, for example, there is a sampling frequency drift or gain modulation that is not appropriately compensated, there is no good overall Least Squares fit, and a significant fraction of the (modulated) MLS signal resides in $\hat{w}(n)$. In fact, parts of the proposed channel model were inspired by residuals observed in $\hat{w}(n)$. We define the estimated signal-to-noise ratio (and error measure) as

$$\mathrm{SNR}_{\mathrm{est}} = 20 \log_{10} \left(\frac{\hat{w}_{\mathrm{RMS}}}{\hat{y}_{H,\,\mathrm{RMS}}} \right) \ \mathrm{dB}.$$

Applying the proposed estimation method to the database using $P = 63$ and always the full data block for input and output, results in the estimated SNRs as summarized in Table 9.2 and the estimated frequency responses as shown in Figure 9.10. A manual inspection of the estimated responses showed that the frequency response mainly depends on the transmission direction (air/ground or ground/air), and as such on the transmitting and receiving voice radios. Apart from this, very little other variation of the frequency response among the different blocks was observed.

Table 9.2 also provides estimation results for two alternative estimation methods. In the 'Least Squares' (LS) method described so far, the full data block is used for the estimation of $h(n)$, which assumes that $h(n)$ is time-invariant. To accommodate for potential slow time-variations of $h(n,t)$, the 'Windowed LS' method performs a Least Squares estimation within a window of five frames (315 samples), and calculates an estimate $\hat{h}(n,t)$ for every sample of the data block. The table also shows the estimation error when tracking $h(n,t)$ with an exponentially weighted recursive-least squares (RLS) adaptive filter [87] with a forgetting factor $\lambda = 0.998$.

The results show that the assumption of a time-invariant $h(n)$ leads to the lowest estimation error. This means that the performance improvement obtained by using as many samples as possible for the estimation (cf. Figure 9.7) outweighs the performance degradation induced by the time-invariance assumption. In other words, the time-variation of the impulse response is smaller than what can be measured given the channel noise. Also, a manual inspection of the error signals did not reveal any significant temporal changes within a data block, which is another justification for the assumption of a time-invariant channel filter $h(n)$.

At low noise levels (e.g., data block number 1), the error signal contains short repetitive noise bursts with a constant frequency of approximately 6 Hz. The origin of these noise bursts is unknown. However, since the same noise bursts are audible in recordings of signal pauses of the database when no MLS signal was transmitted, we conclude that the bursts are not an estimation error but are present in the signal. Due to their periodic occurrence it appears likely that the bursts result from electromagnetic interference within the measurement system itself or between the measurement system and other components of the aircraft or ground systems. At medium noise levels

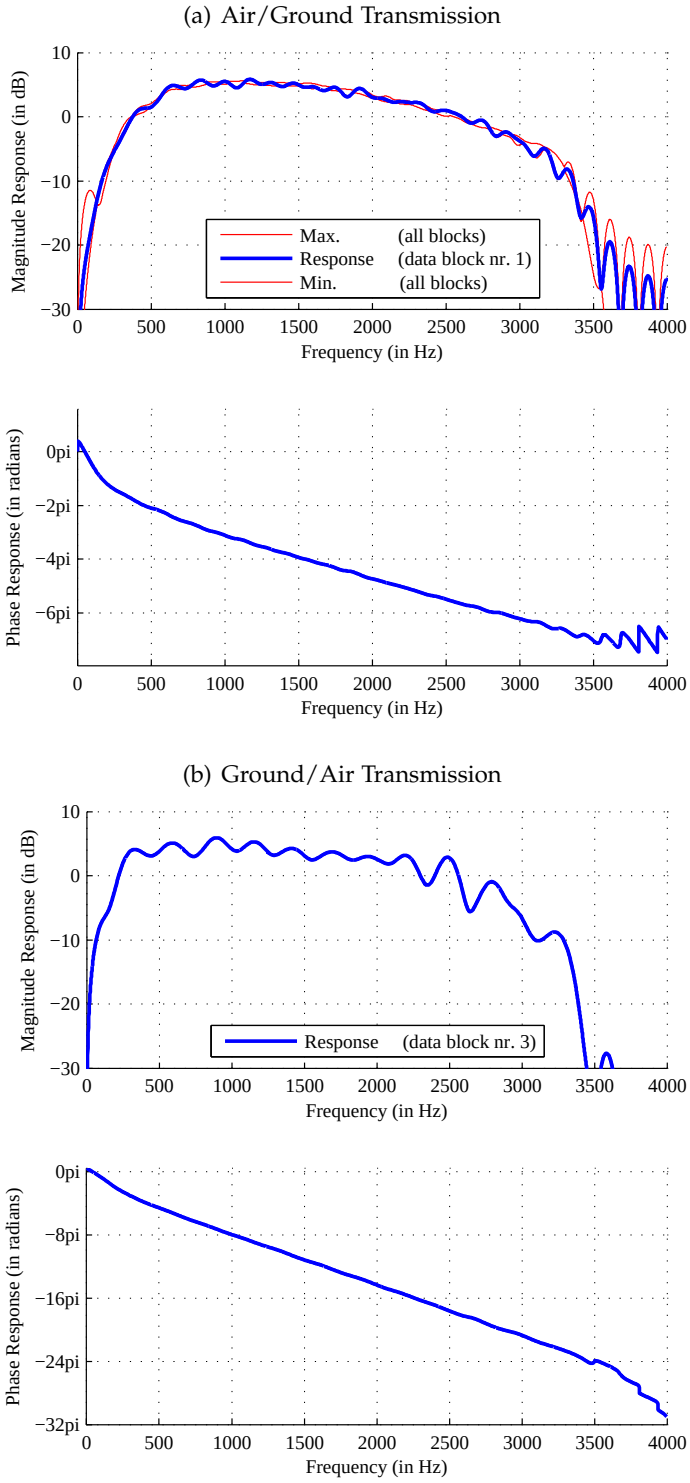
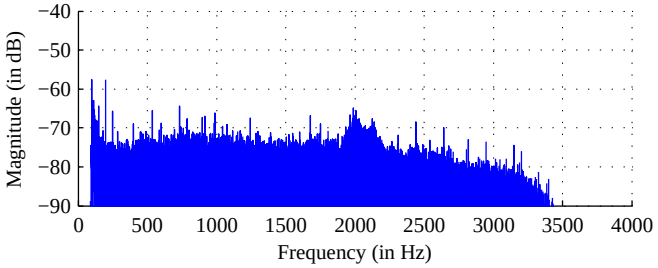


Figure 9.10.: Estimated frequency responses.

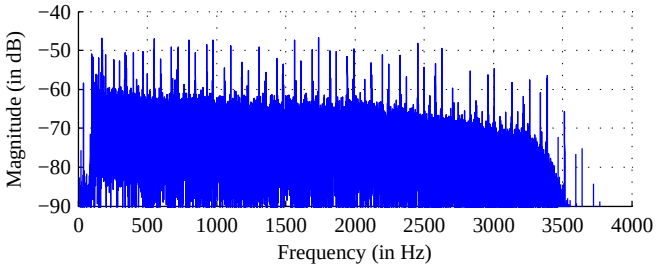
Table 9.2.: Filter and Noise Estimation Results

Data Block			Experimental Conditions					$f_2 - f_0$	SNR_{est} (in dB)			Comment
Number	Frame-ID (Start)	Frame-ID (End)	Transm. Direction	Engine On/Off	Distance (in km)	Speed (in m/s)	Altitude (in m)	Rel. Azimuth (in °)	Difference (in Hz)	Least Squares	Windowed LS	Recursive LS	
1	2012056	2013560	A/G	Off	0.1	0.0	297	n/a	0.96	39.5	38.5	37.0	Back-to-Back
2	2065388	2066892	A/G	On	0.1	2.2	296	n/a	0.94	36.7	35.8	36.0	Back-to-Back
3	1008000	1009504	G/A	Off	0.1	0.0	300	n/a	-0.09	37.6	36.7	36.6	Back-to-Back
4	4016698	4018202	G/A	On	0.1	0.3	295	n/a	-0.10	35.3	34.1	34.6	Back-to-Back
5	2396236	2397740	A/G	On	5.9	61.7	1467	166	0.92	28.8	28.1	28.1	Pos. Doppler Shift
6	2400075	2401579	A/G	On	4.0	64.2	1342	173	0.92	27.4	26.6	27.0	Pos. Doppler Shift
7	2354363	2355867	A/G	On	5.8	58.2	1363	197	0.92	27.5	26.7	27.0	Pos. Doppler Shift
8	2412064	2413568	A/G	On	2.2	51.9	1224	26	0.94	22.9	22.3	22.8	Neg. Doppler Shift
9	2377571	2379075	A/G	On	3.9	41.5	1343	331	0.93	29.3	28.3	28.9	Neg. Doppler Shift
10	2116191	2117695	A/G	On	5.7	40.6	1032	3	0.89	30.8	29.7	30.3	Neg. Doppler Shift
11	2251911	2253415	A/G	On	7.3	39.2	1594	268	0.98	22.7	22.3	22.4	Zero Doppler Shift
12	4260506	4262010	G/A	On	19.1	58.0	992	183	-0.10	30.5	29.9	29.9	Med. Distance
13	3292995	3294499	A/G	On	43.5	50.1	1075	92	0.99	23.0	22.1	22.8	Max. Distance

(a) Data Block Nr. 1 (A/G, engine off, back-to-back)



(b) Data Block Nr. 7 (A/G, engine on, flight)



(c) Data Block Nr. 13 (A/G, engine on, flight, max. distance)

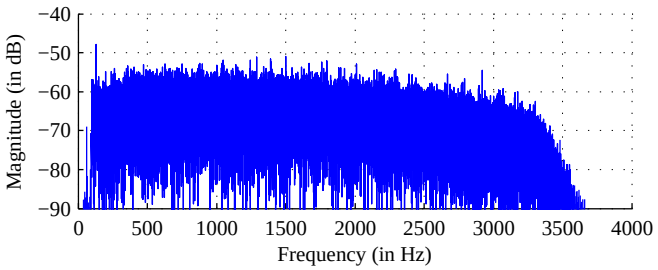


Figure 9.11.: DFT magnitude spectrum of the noise estimate and estimation error signal $\hat{w}$ for different data blocks.

(e.g., data block number 7) the error signal sounds similar to colored noise, and at the highest noise levels (e.g., data block number 13), the error signal is very similar to white noise. Figure 9.11 shows the corresponding power spectra of the error signals.

9.4.2. DC Offset

Figure 9.12 shows the time-variant DC offset $c(n)$ in the recordings of the received signal, obtained by filtering data block number 1 with a lowpass filter of order 1000 and a cutoff frequency of 40 Hz. The clearly visible pulses with a frequency of 1 Hz result from interference between the voice recordings and the GPS timing signal within the measurement system, and are as such a measurement error. Even though the amplitude of this DC offset is small compared to the overall signal amplitude, the offset, if not

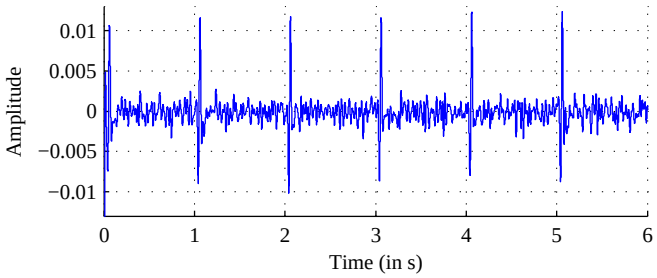


Figure 9.12.: DC component of the received signal recording. The pulses result from interference with the GPS timing signal.

removed a-priori, would severely degrade the overall estimation performance.

9.4.3. Gain

In some data blocks we observed a sinusoidal amplitude modulation of the estimated gain $\hat{g}$ of the received signal. The DFT magnitude spectrum of the time-variant gain $\hat{g}(n)$ for three different data blocks is shown in Figure 9.13. In two of these cases the spectrum exhibits a distinct peak. For every data block, Table 9.3 provides the frequency location $f_{\sin}$ of the most dominant spectral peak, as well as its amplitude $A_{\sin}$ relative to the DC gain C, expressed in

$$L_{A_{\sin}/C} = 20\log_{10}\left(\frac{A_{\sin}}{C}\right),$$

modeling the gain modulation with $g(n) = C + A_{\sin}\sin\left(\frac{2\pi f_{\sin}}{f_s}n\right)$. Expressing the same quantity in terms of RMS values results in

$$L_{A_{\mathrm{RMS}}/C_{\mathrm{RMS}}} = 20\log_{10}\left(\frac{\left[A_{\sin}\sin\left(\frac{2\pi f_{\sin}}{f_s}\right)\right]_{\mathrm{RMS}}}{C_{\mathrm{RMS}}}\right) \approx L_{A_{\sin}/C} - 3\,\mathrm{dB}.$$

Table 9.3 also shows the overall power of the gain modulations relative to the DC gain, expressed in

$$L_{g-C/C} = 20\log_{10}\left(\frac{[\hat{g}(n) - C]_{\mathrm{RMS}}}{C_{\mathrm{RMS}}}\right).$$

The source of the sinusoidal gain modulations (GM) is not immediately clear. Given Figure 9.13 and Table 9.3, we observe the following facts:

1. GM is present if engine is running, only.
2. GM is present even if aircraft is not moving.
3. Frequency of GM is independent of the relative heading of the aircraft.

4. Frequency of GM depends on an unknown variable that has some connection with the aircraft speed.

5. Frequency of GM is significantly larger than the frequency predicted in Appendix B for gain modulations caused by multipath propagation and Doppler shift.

We conclude from these facts that the observed gain modulation is not caused by the physical radio transmission channel, especially since it as also present in situations where the aircraft did not move.

We believe that the gain modulation results from interference between the aircraft power system and the voice radio or the measurement system. In particular, the gain modulation might result from a ripple voltage on the direct current (DC) power supply of the aircraft radio. Such a ripple voltage is a commonly observed phenomenon and results from insufficient filtering or voltage regulation of the rectified alternating current (AC) output of the aircraft's alternator or power generator [149]. The frequency of the ripple voltage is, as such, a linear function of the engine's rotational speed, which is measured in rotations per minute (rpm). We conjecture that the unknown variable that the gain modulation frequency depends on, is the engine rpm. This would align well with the results of Table 9.3 with no gain modulation when the engine is off and a significant frequency difference between engine idle and engine full throttle. The magnitude of the ripple voltage as well as its impact on the communication system is expected to be system- and aircraft-specific and to be less of an issue in modern systems and large jet aircraft with an independent auxiliary power unit (APU).

An alternative explanation for the gain modulation could be a periodic variation of the physical radio transmission channel induced by the rotation of the aircraft's propeller. The rotational speed of the propeller is linked to the engine rpm either directly or via a fixed gear transmission ratio.

9.5. Conclusions

In this section, we presented an improved model to describe the data in the *TUG-EEC-Channels* database presented in Chapter 8 and a corresponding estimation method. This method allows to characterize the measured channel in terms of its linear filter response, time-variant gain, sampling offset, DC offset and additive noise, and also helps to identify measurement errors. In contrast, the channel estimates included in the database combine all effects into a single variable, namely a channel response estimated frame by frame.

Given the low level of the estimation error signal and the spectral characteristics of the noise estimate, we conclude that the proposed estimation method works as expected, and that the proposed data model is a reasonable description for the measured data. As expected by theory, the analysis of the measurements confirms that the channel is not frequency-selective. The source of the observed flat fading could not be conclusively established. However, there are several factors that strongly suggest that the observed gain modulation does not result from radio channel fading through

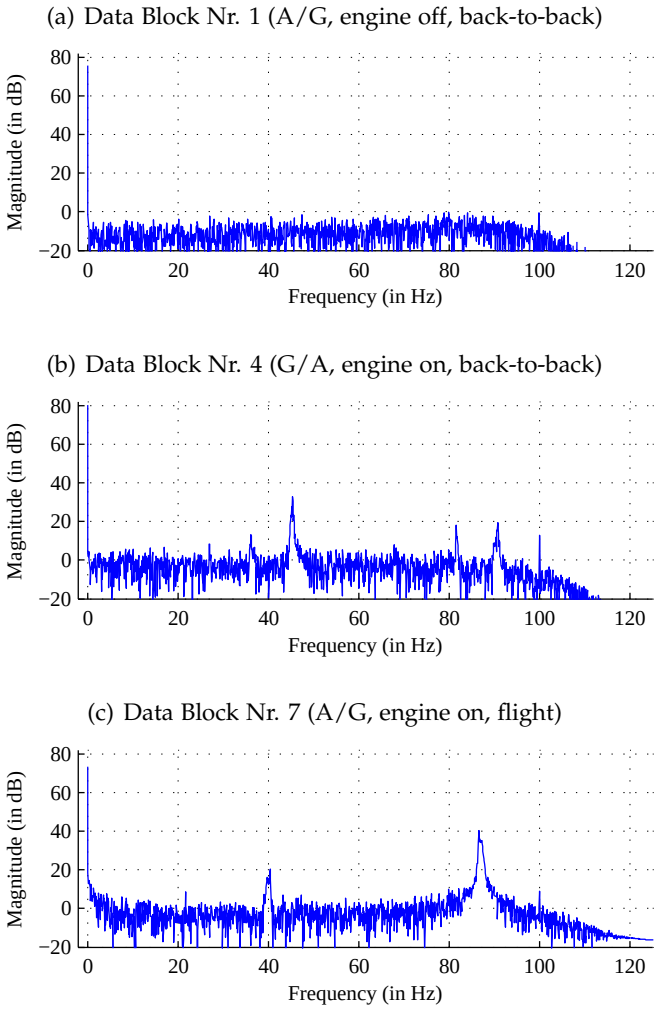


Figure 9.13.: DFT magnitude spectrum of the time-variant gain $\hat{g}$ of different data blocks.

Table 9.3.: Dominant Frequency and Amplitude Level of the Gain Modulation

Data Block Nr.	Frequency	$L_{A_{\sin}/C}$	$L_{A_{\mathrm{RMS}}/C_{\mathrm{RMS}}}$	$L_{g-C/C}$
1	79.9 Hz	-69.7 dB	-72.8 dB	-51.2 dB
2	43.2 Hz	-53.2 dB	-56.3 dB	-42.4 dB
3	27.1 Hz	-64.8 dB	-67.9 dB	-51.1 dB
4	45.4 Hz	-41.4 dB	-44.5 dB	-39.3 dB
5	86.8 Hz	-28.3 dB	-31.3 dB	-23.3 dB
6	86.5 Hz	-27.6 dB	-30.6 dB	-21.9 dB
7	86.6 Hz	-26.8 dB	-29.8 dB	-22.9 dB
8	83.7 Hz	-27.3 dB	-30.3 dB	-25.1 dB
9	81.8 Hz	-36.7 dB	-39.8 dB	-34.8 dB
10	80.2 Hz	-34.8 dB	-37.8 dB	-36.0 dB
11	73.0 Hz	-24.2 dB	-27.2 dB	-21.8 dB
12	86.5 Hz	-33.1 dB	-36.1 dB	-31.6 dB
13	65.1 Hz	-36.6 dB	-39.6 dB	-31.7 dB

multipath propagation but from interference between the aircraft engine and the voice radio system. The observed noise levels are in a range from 40 dB to 23 dB in terms of SNR, which are worst-case estimations since all estimation errors accumulate in the noise estimate, and the real channel noise level might in fact be smaller.

Chapter 10

Experimental Watermark Robustness Evaluation

In this chapter, we make use of the channel model derived in Chapter 9 to evaluate the robustness of the proposed watermarking method in the aeronautical application. We experimentally demonstrate the robustness of the method against filtering, desynchronization, gain modulation and additive noise. Furthermore we show that pre-processing of the speech signal with a dynamic range controller can improve the watermark robustness as well as the intelligibility of the received speech.

This chapter presents recent results that have not yet been published.

This chapter experimentally evaluates the robustness of the watermark system presented in Chapter 4 in light of the application to the air traffic control voice radio. The evaluation is an explicit continuation of the experimental results of Section 4.3, now incorporating the obtained knowledge about the aeronautical voice radio channel. Consequently, we use the same experimental setup and settings as described in Section 4.3.

10.1. Filtering Robustness

Section 4.3 demonstrated the robustness of the watermarking method against linear and time-invariant filtering using a bandpass filter, an IRS filter, a randomly chosen estimated aeronautical channel response filter and an allpass filter. These results are repeated and extended in Figure 10.1.

10.1.1. Estimated Static Prototype Filters

We evaluated the filtering robustness using the measured and estimated aeronautical radio channel filters of Chapter 9. In particular, we used the two distinct channel response filters for air/ground and ground/air transmissions as depicted in Figure 9.10.

The resulting BERs are shown in Figure 10.1 and indicate that the method is robust against both filters (denoted 'Ch. 9, air/gnd' and 'Ch. 9, gnd/air'). The increased BER in the ground/air transmission direction likely results from a mismatch between the used embedding band position (666 Hz–3333 Hz) and the frequency response of the ground/air channel. As shown in Figure 9.10, the ground/air channel has a non-flat frequency response and significant attenuation above 2.5 kHz. A more suitable choice for the watermark embedding would be the configuration described in Section 4.2.2.3 with an embedding band from 0.5 kHz to 2.5 kHz.

10.1.2. Time-Variant *TUG-EEC-Channels* Filter

We evaluated the robustness against the time-variant CPSD-based channel estimates of Section 8.2.5.4. This is a worst-case scenario, since, as discussed in Chapter 9, the time-variation of these channel response estimates originates from an estimation error due to insufficient compensation of the sampling clock drift.

We used the estimates of two different measurement scenarios, one with the aircraft being static on ground ('Ch. 8, ground', Frame-ID 2011556 to 2013968 in the *TUG-EEC-Channels* database) and one with aircraft flying at high speed ('Ch. 8, flight', Frame-ID 2396236 to 2399927), both with transmission in air/ground direction. The filter taps obtained from the CPSD-based channel estimates (available every 7.875 ms) were interpolated over time and continuously updated while filtering the watermarked speech signal. The results in Figure 10.1 confirm the robustness of the watermarking method against this time-variant filtering attack.

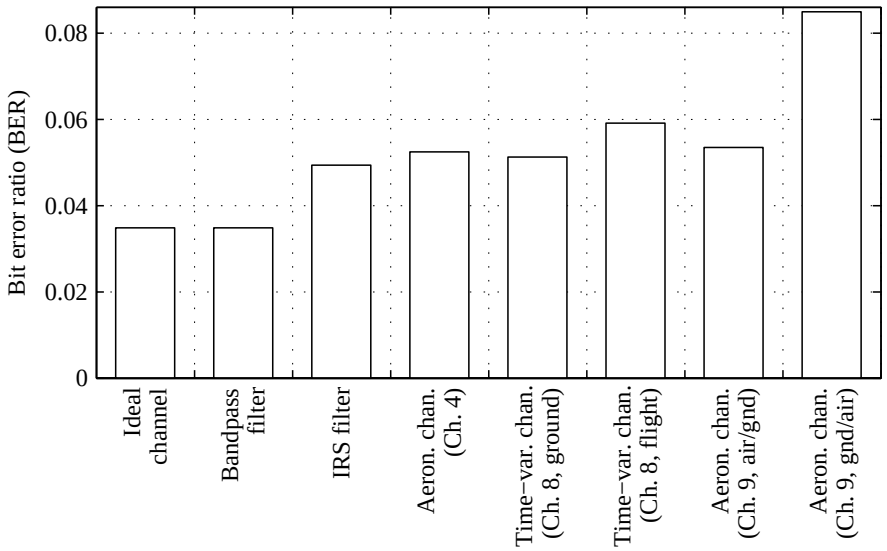


Figure 10.1.: Overall system robustness in the presence of various transmission channel filters at an average uncoded bit rate of 690 bit/s.

10.2. Gain Modulation Robustness

In this section, the robustness of the watermarking method against sinusoidal gain modulations is evaluated, both systematically and in the measured aeronautical channel conditions.

In contrast to the previous experiments, in this and all following experiments of this chapter the input speech is normalized to unit variance on a per utterance level to establish a coherent signal level across utterances.

10.2.1. Sinusoidal Gain Modulation Robustness

Figure 10.2 shows the robustness of the principle embedding method against sinusoidal gain modulations of the form

$$\hat{s}_{\mathrm{FB}}(n) = \left[1 + h_{\mathrm{GM}} \sin\left(2\pi n \tfrac{f_{\mathrm{GM}}}{f_s}\right)\right] s'_{\mathrm{FB}}(n)$$

with a modulation frequency f_{GM} and a modulation depth h_{GM}. The experimental conditions are the same as in Section 4.3, except that ideal frame synchronization is assumed since the implemented frame detection scheme works reliably only up to a BER of approximately 15 % (see Figure 5.8).

Using the measured worst-case modulation frequency and depth of Table 9.3 (Data Block Nr. 11, $f_{\mathrm{GM}} = 73\,\mathrm{Hz}$, $h_{\mathrm{GM}} = -24.2\,\mathrm{dB} = 0.062$) and real frame detection, the resulting BER is shown in comparison to an ideal channel with constant gain in Figure 10.3.

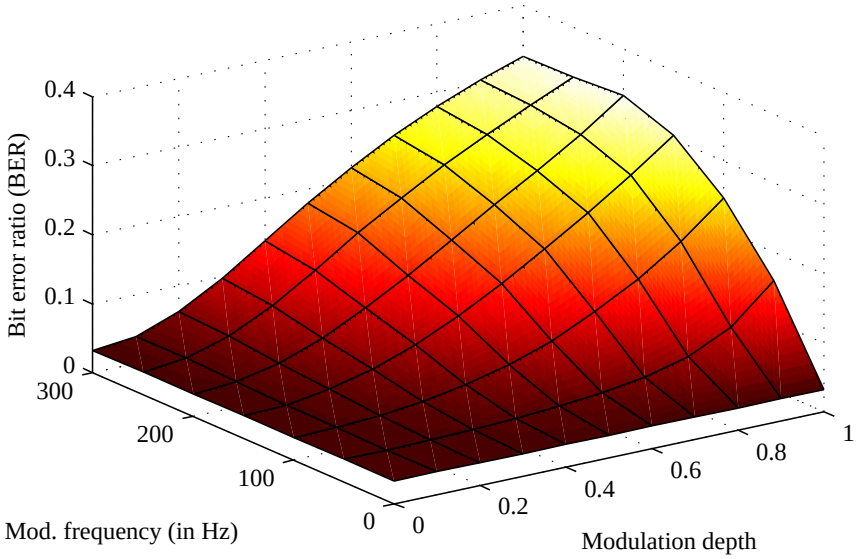


Figure 10.2.: Watermark embedding scheme robustness against sinusoidal gain modulation at an average uncoded bit rate of 690 bit/s (assuming ideal frame synchronization).

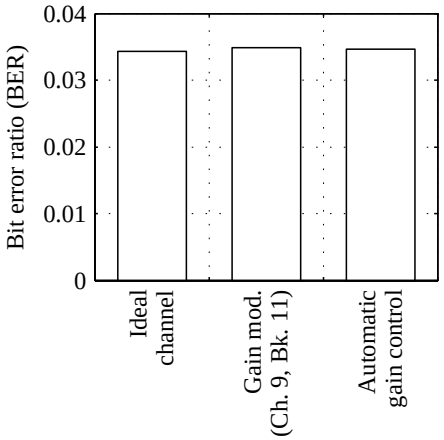


Figure 10.3.: Overall system robustness in the presence of the measured gain modulation and a simulated automatic gain control at an average uncoded bit rate of 690 bit/s (and using the actual frame synchronization).

10.2.2. Simulated Automatic Gain Control

Aeronautical radio transceivers typically contain an automatic gain control (AGC) both in the radio frequency (RF) and the audio domain [97, 150]. To evaluate the robustness against adaptive gain modulation, we simulated a channel with an AGC using the SoX implementation [151] of a dynamic range controller (or 'compressor', [152]) with an attack time of 20 ms, a release time of 500 ms, a look-ahead of 20 ms, and a compression ratio of 1:4 above a threshold level of -24 dB [150]. The resulting BER shown in Figure 10.3 demonstrates that the AGC does not decrease the watermark detection performance.

10.3. Desynchronization Robustness

An experimental evaluation of the timing, bit and frame synchronization is provided in Section 5.3. This topic is not further treated herein.

10.4. Noise Robustness

The robustness of the watermarking method against AWGN was evaluated in Section 4.3 with a segmental SNR of 30 dB. The noise robustness for a wide range of non-segmental SNRs is shown in Figure 10.4 ('Original signal'). To show the behavior of the principle method, perfect frame detection is assumed, since otherwise the BER would level off at around 15 % due to failing frame synchronization.

The BER curve indicates that the method would sometimes not be robust against the measured aeronautical channel conditions, since the worst-case measured SNR (including estimation errors that contribute to the noise level) was found to be 22.7 dB (see Table 9.2). However, the robustness of the watermarking scheme against AWGN is difficult to specify in absolute terms in the aeronautical application. Due to the presence of gain controls in the transceivers, the transmission channel is non-linear, and care must be taken when defining or comparing signal-to-noise ratios. While SNR is a measure of the mean signal and noise power, the aeronautical channel is in fact peak power constrained, with the peak power level being given by the maximum allowed output power of the transmitter (integrated over a short time window). The BER curves corresponding to Figure 10.4 are shown in Figure C.1 of Appendix C for given *peak* signal-to-noise ratios (PSNR). We assume the peak power to be the signal's maximum average power measured in windows of 20 ms.

The SNRs of Table 9.2 were measured with an input signal of constant instantaneous power. As a consequence, the automatic gain controls were in a steady state and the output signal power at some nominal level. Due to the reaction time of the gain controls, a peak of a real speech signal may in practice far exceed this nominal level, and the actual peak power limit lies above this measured steady-state level.

In the remainder of this section, we briefly discuss and experimentally evaluate additional measures to increase the noise robustness of the watermarking method.

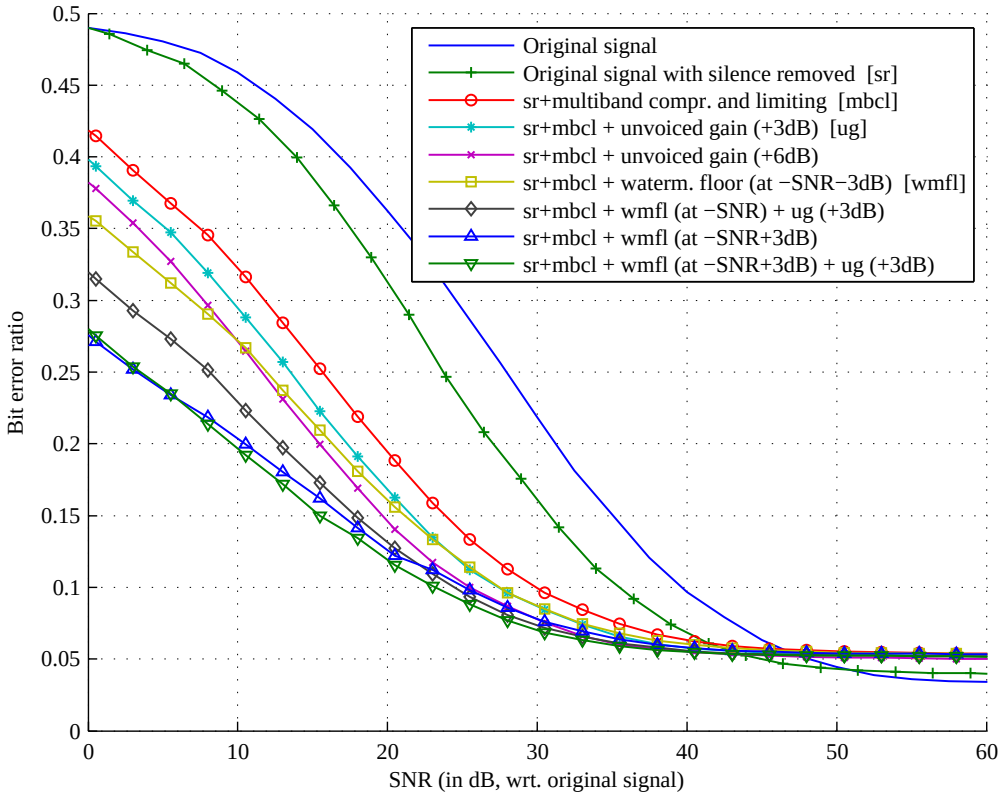


Figure 10.4.: Watermark embedding scheme robustness against AWGN at average uncoded bit rates of 690 bit/s ('original signal'), 470 bit/s ('sr') and 485 bit/s ('sr+mbcl').

10.4.1. Selective Embedding

The watermarking method as described in Chapter 4 embeds the watermark in all time segments that are not voiced, which also includes pauses and silent segments. However, when assuming a fixed channel noise level, the local SNR in silent regions is very low and watermark detection fails, leading to high BERs. A simple measure to increase the noise robustness of the watermarking method is to not embed during pauses and constrain the embedding to regions where the speech signal has significant power. Inevitably, this decreases the uncoded watermark bit rate.

We simulate the non-embedding in pauses by removing all silent regions of the input signal, which are defined by a 43ms-Kaiser-window-filtered intensity curve being more than 30 dB below its maximum for longer than 100ms. Given the ten normalized utterances of the ATCOSIM corpus used throughout the tests of this chapter, 14 % of the speech signal are detected as silence. This reduces the average watermark bit rate from 690 bit/s to 470 bit/s, both calculated on the basis of the duration of the original signal. The resulting BER curve shown in Figure 10.4 ('silence removal') demonstrates a significant improvement in terms of noise robustness, however, at the cost of a reduced

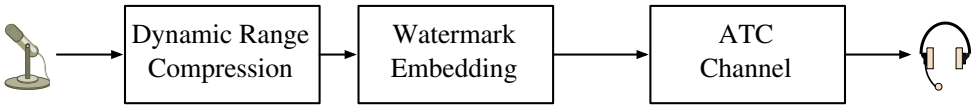


Figure 10.5.: Improving watermark robustness and speech intelligibility by dynamic range compression.

bit rate.

10.4.2. Speech Enhancement and Watermarking

A Master's thesis [131] that was supervised by the author of this work studied and compared different algorithms to pre-process the speech signal before the transmission over the aeronautical radio channel in order to improve the intelligibility of the corrupted received speech. The compared systems are fully backward-compatible and operate only at the transmitting side. The methods were compared subjectively using the modified rhyme test (MRT) and significantly improved the intelligibility of the received speech at high channel noise levels.

The study showed the high effectiveness of a dynamic range controller (or 'compressor', [152]) to improve the intelligibility. The isolated word recognition rate increased by 12.6 percentage points on average. Given the peak power constrained transmission channel and, thus, a fixed peak signal to noise level ratio, the compressed speech signal has much higher overall energy compared to the original signal and, consequently, a larger perceptual distance from the constant-power channel noise.

The compressor increases the SNR of the transmitted signal without increasing the peak signal-to-noise ratio (PSNR). As a consequence, the dynamic range compression not only increases the intelligibility, but also the noise robustness of the watermark system. Since the speech signal has overall higher energy, also the watermark signal that replaces the speech signal is higher in energy relative to the channel noise if the compression precedes the watermark embedding as shown in Figure 10.5.

We simulate a dynamic rage controller using Audacity [153] with the Apple CoreAudio implementation [154] of a multiband compressor followed by the LASDPA implementation [155] of a limiter [152]. The settings used for both units were chosen manually and are shown in Figure C.2 of Appendix C.

The input speech signal without silent regions as used in the previous subsection was pre-processed by the compressor before the embedding of the watermark as shown in Figure 10.5. Given various channel noise levels, the resulting BER curve is shown in Figure 10.4 and exhibits a significant improvement.

Due to the presence of the compression system, it is of relevance if the signal to channel noise ratio is defined as SNR or PSNR. Further on, the SNR is different depending on whether the signal power is measured before the enhancement system (as the channel is peak-power constrained and both processed and unprocessed signal have the same peak power) or if it is measured at the channel input. Figure C.1 shows the BER curves corresponding to Figure 10.4 but using alternative signal-to-noise ratio

measures.

We cannot overstate the fact that within a single system that combines watermark embedding and compression it is possible to robustly embed a watermark and, at the same time, increase the intelligibility of the transmitted ATC speech. The system could be fully backward-compatible. At the receiving end all receivers (including legacy installations) would benefit from the intelligibility improvement. An equipped receiver is required in order to benefit also from the embedded watermark.

10.4.3. Watermark Amplification

In principle, any measure that increases the power of the embedded watermark signal relative to the channel noise increases the likelihood that the embedded message will be successfully detected. Two simple measures are introduced and their effectiveness evaluated.

A primitive way to increase the watermark power is to add the watermark data signal $w(n)$ (using the notation of Section 4.1) to the watermarked signal $\hat{s}_{\mathrm{PB}}(n)$. The watermark signal is added as a watermark floor at a fixed power level relative to the channel noise power. In non-voiced regions, a comfort noise of equivalent power is added in order to maintain a constant background noise level. The resulting BER curves shown in Figure 10.4 ('waterm. floor') demonstrate the effectiveness of the watermark floor. Given a watermark floor level -3 dB below the channel noise, the watermark floor is imperceptible since it is masked by the channel noise. Nevertheless it significantly improves the noise robustness of the watermark. At higher levels, the watermark floor becomes audible, and the additional gain in robustness comes at the cost of perceptual quality. If the channel noise level is not known a-priori, a certain value has to assumed based on operational requirements.

Another simple measure to increase the watermark power is to increase the gain $g(n)$ of the non-voiced speech components (cf. Section 4.1). Figure 10.4 ('unvoiced gain') shows the positive effect on the BER curves given a gain $g(n)$ that is increased by a factor of +3 dB and +6 dB relative to its initial measured value. Since the non-voiced components are often low in power compared to voiced components, the additional gain does in most cases not increase the peak power of the watermarked signal. The increased noise robustness comes, however, at the cost of perceptual quality.

10.5. Conclusions

We conclude that the proposed watermarking method is robust against any filtering that is expected to occur on the aeronautical radio channel. The method allows the embedding of the watermark in a frequency band that matches the passband of the transmission channel. The equalizer in the watermark detector compensates for phase distortions and tracks eventual time-variations of the channel filters. The high robustness against filtering attacks comes as no surprise since the watermark detector also has to cope with the rapidly time-varying LP synthesis (vocal tract) filter.

The watermarking method is also highly robust against sinusoidal gain modulations

over a wide range of modulation frequencies and modulation depths, considering that the measured worst case modulation depth is well below -20 dB. Non-sinusoidal gain modulations in the form of an AGC do not degrade the detection performance either. The high gain modulation robustness is not surprising because the method must inherently be robust against the rapidly time-varying gain of the unvoiced speech excitation signal.

The timing synchronization scheme based on the detector's equalizer shows satisfactory performance, but there is room for further improvement by the application of fractionally spaced instead of sample-spaced equalizers. A working implementation for frame synchronization was presented, but many alternative schemes exist and further optimization is possible.

We evaluated the noise robustness over a wide range of signal-to-noise ratios. It is difficult to specify the SNR ranges of real world ATC channels. The implementation presented herein far exceeds the operationally required watermark bit rate of approximately 100 bit/s [116]. Given the presented BER curves, for an operational ATC implementation it might be beneficial to reduce the bit rate and, therefore, increase the noise robustness of the watermark.

A number of further measures are available to improve the noise robustness. The most remarkable is a pre-processing of the speech signal by multiband dynamic range compression and limiting. This processing not only boosts the noise robustness of the watermark, but also increases the intelligibility of the received speech signal.

Chapter 11

Conclusions

Air traffic control (ATC) radio communication between pilots and controllers is subject to a number of limitations given by the legacy system currently in use. To overcome these limitations, and ultimately increase safety, security and efficiency in ATC, this thesis investigates the embedding of digital side information into ATC radio speech. It presents a number of contributions towards the ATC domain as well as the area of robust speech watermarking. Figure 11.1 illustrates the connections among the different contributions in the context of the ATC application at hand. The structure of this chapter mostly mirrors the signal flow shown in the figure.

A review of related work reveals that most of the existing watermarking theory and most practical schemes are based on a non-removability requirement, where the watermark should be robust against tampering by an informed and hostile attacker. While this is a key requirement in, e.g., transaction tracking applications, it poses an unnecessary constraint in the legacy enhancement application considered in this thesis. Theoretical capacity derivations show that dropping this constraint allows for significantly larger watermark capacities, because watermarking can be performed in perceptually irrelevant instead of perceptually relevant host signal components. In contrast to perceptually relevant components, perceptually irrelevant signal components are not limited to small modifications but can be replaced by arbitrary values.

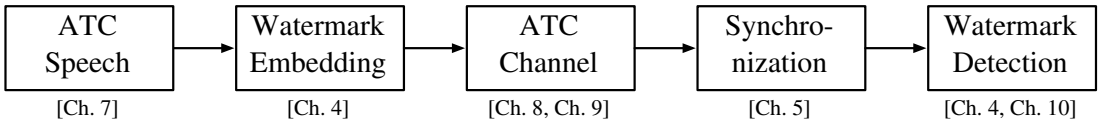


Figure 11.1.: Relationship among the different chapters of this thesis.

We derived the watermark capacity in speech based on perceptual masking and the auditory system's insensitivity to the phase of unvoiced speech, and showed that this capacity far exceeds the capacity of the conventional ideal Costa scheme.

To underline the validity of our theoretical results, we designed and implemented a practical speech watermarking method that replaces the imperceptible phase of non-voiced speech by a watermark data signal. Even though the practical scheme still operates far below the theoretical capacity limit, it outperforms most current state-of-the-art methods and shows high robustness against a wide range of signal processing attacks.

To validate the robustness of our scheme in the aeronautical application, we created two evaluation resources. At first, good knowledge about the characteristics of the aeronautical voice radio channel is required. Since little information is publicly available, we designed and implemented a ground/airborne based audio channel measurement system and performed extensive in-flight channel measurements. The collected data is now publicly available and can be used as a basis for further developments related to the air/ground voice radio channel. We proposed a model for the measured channel data, and the extracted model parameters are used in the validation of the watermarking scheme to simulate different effects of the aeronautical channel.

In addition, due to the particular nature of ATC speech and the limited availability of ATC language resources, we also created an ATC speech corpus. The corpus is useful in the validation of our watermarking scheme, but also constitutes a valuable resource for language research and spoken language technologies in ATC.[1]

Synchronization between watermark embedder and detector is necessary to detect the watermark in the received signal. Many proposed watermarking schemes assume perfect synchronization and do not treat this (often difficult) problem any further. We carefully addressed this issue, transferred digital communication synchronization methods to the watermarking domain, and presented a full implementation that covers the various layers of synchronization.

The watermark detector is based on hidden training sequences and adaptive equalization. The overall method is inherently robust against time-variant filtering and gain modulation, which is a distinct advantage compared to quantization-based watermarking. Moreover, the proposed scheme outperforms current state-of-the-art methods on a scale of embedded bit rate per Hz of host signal bandwidth, considering the type of channel attacks that are relevant in the aeronautical application.

Using the data obtained in the channel measurements, we evaluated the robustness of the proposed method in light of the aeronautical application. The method is robust against static and time-variant measured channel filter responses, the measured channel gain modulation, and a simulated automatic gain control. Due to the non-linearity of the ATC channel it is difficult to specify absolute channel noise levels. We evaluated the noise robustness over a large range of channel SNRs and proposed a number of measures that, if necessary, can increase the noise robustness of the proposed

[1] Within one year since its public release, the speech corpus DVD with a download size of 2.4 GB was requested by mail from several institutions, and was downloaded in full from our website by approximately 40 people (conservative estimate based on manually filtered server log-files).

method. We recommend a pre-processing of the input speech signal with multiband dynamic range compression and limiting for the aeronautical application. This not only significantly increases the noise robustness of the watermark but also improves the subjective intelligibility of the received speech. An increased intelligibility in air/ground voice communication would constitute a further substantial contribution to the safety in air traffic control.

Naturally, the work presented in this thesis is subject to limitations, and a number of options for further research arise.

Since we proposed an entirely novel watermarking scheme, our implementation can be considered as proof-of-concept implementation, only. The large gap to the theoretical capacity shows that the performance can be further increased and numerous options for further optimizations exist. For example, it is expected that an optimization of the existing parameters, an SNR-adaptive watermark embedding, fractionally-spaced equalization, or a joint synchronization, equalization and detection scheme could increase both robustness and data rate. Furthermore, we have not evaluated the real-time capability of our implementation. While the method operates in principle on a per frame basis and is computationally not prohibitively expensive, a real-time implementation was outside the scope of this thesis and is yet to be carried out. The evaluation of the method's robustness in the aeronautical application was limited by the relatively small scope of the available channel data. To improve the validity of the evaluation, the aeronautical channel measurements should be carried out on a large scale with a wide range of aircraft types, transceiver combinations and flight situations.

We expect that a continuation of the theoretical work laid out in this thesis could lead to a deeper understanding of speech watermarking and, ultimately, to better performing system designs. Our theoretical capacity estimation for phase modulation based watermarking is an upper bound, only, due to the idealistic assumptions made about the host signal and the transmission channel. An evaluation of the capacity under tighter but more realistic assumptions would provide additional insight and likely result in a significantly lower watermark capacity. Complementing our experimental robustness evaluation, a theoretical analysis of the robustness of the phase modulation method against different transmission channel attacks could provide additional insight. Incorporating the particular characteristics of the host signal and the aeronautical channel, such an analysis would likely lead to a better system design. Last but not least, the capacity derivations for masking based watermarking show that in our system a lot of additional watermark capacity is still unused because perceptual masking is not incorporated. Further research in speech watermarking should aim to exploit this additional capacity by combining masking-based principles with the phase replacement approach introduced in this thesis.

Appendix A

ATCOSIM Transcription Format Specification

This appendix provides the formal specification of the transcription format of the ATCOSIM speech corpus presented in Chapter 7. While Section A.1 describes the basic rules of the transcription format, Section A.2 contains content-specific amendments such as lists of non-standard vocabulary.

A.1. Transcription Format

The orthographic transcription follows a strict set of rules which is presented hereafter. In general, all utterances are transcribed word-for-word in standard British English. All standard text is written in lower-case. Punctuation marks including periods, commas and hyphens are omitted. Apostrophes are used only for possessives (e.g. `pilot's radio`)[1] and for standard English contractions (e.g. `it's`, `don't`).

Technical noises as well as speech and noises in the background—produced by speakers other than the one recorded—are not transcribed. Silent pauses both between and within words are not transcribed either. Numbers, letters, navigational aids and radio call signs are transcribed as follows.

In terms of notation, stand-alone technical mark-up tags are written in upper case letters with enclosing squared brackets (e.g. `[HNOISE]`). Regular lower-case letters and words are preceded or followed by special characters to mark truncations (`=`), individually pronounced letters (`~`) or unconfirmed airline names (`@`). Groups of words are embraced by opening and closing XML-style tags to mark off-talk (`<OT>`

[1] The `mono-spaced typewriter` type represents words or items that are part of the corpus transcription.

... </OT>), which is also transcribed, and foreign language (<FL> </FL>), for which currently no transcription is given.

A.1.1. ICAO Phonetic Spelling

Definition: Letter sequences that are spelled phonetically.

Notation: As defined in the reference.

Example: `india sierra alfa report your heading`

Word List: `alfa bravo charlie delta echo foxtrot golf hotel india juliett kilo lima mike november oscar papa quebec romeo sierra tango uniform victor whiskey xray yankee zulu`

Supplement: `fox` (short form for `foxtrot`)

Reference: [113]

A.1.2. Acronyms

Definition: Acronyms that are pronounced as a sequence of separate letters in standard English.

Notation: Individual lower-case letters with each letter being preceded by a tilde (~). The tilde itself is preceded by a space.

Example: `~k ~l ~m`

Exception: Standard acronyms which are pronounced as a single word. These are transcribed without any special markup.

Exception Example: NATO, OPEC, ICAO are transcribed as `nato`, `opec` and `icao` respectively.

A.1.3. Numbers

Definition: All digits, connected digits, and the keywords 'hundred', 'thousand' and 'decimal'.

Notation: Standard dictionary spelling without hyphens. It should be noted that controllers are supposed to use non-standard pronunciations for certain digits, such as 'tree' instead of 'three', 'niner' instead of 'nine', or 'tousand' instead of 'thousand' [113]. This is however applied inconsistently, and in any case transcribed with the standard dictionary spelling of the digit.

Example: `three hundred`, `one forty four`, `four seven eight`, `one oh nine`

Word List: `zero oh one two three four five six seven eight nine ten hundred thousand decimal`

A.1.4. Airline Telephony Designators

Definition: The official airline radio call sign.

Notation: Spelling exactly as given in the references.

Examples: `air berlin`, `britannia`, `hapag lloyd`

Exceptions: Airline designators given letter-by-letter using ICAO phonetic spelling as well as airline designators articulated as acronyms.

Exceptions Examples: `foxtrot sierra india`, `~k ~l ~m`

References: [156, 157, 129]

A.1.5. Navigational Aids and Airports

Definition: Airports and navigational aids (navaids) corresponding to geographic locations.

Notation: Geographical locations (navaids) are transcribed as given in the references using lower-case letters. The words used can be names of real places (ex. `hochwald`) or artificial five-letter navaid or waypoint designators (ex. `corna`, `gotil`).[2] Airports and control centers are transcribed directly as said and in lower-case spelling.

Examples:
```
contact rhein on one two seven
alitalia two nine two turn left to gotil
alitalia two nine two proceed direct to corna charlie oscar romeo november
alfa
```

References: [158, Annex A: Maps of Simulation Airspace], [159, 160]

A.1.6. Human Noises

Definition: Human noises such as coughing, laughing and sighs produced by the speaker. Also breathing noises that were considered by the transcriptionist as exceptionally loud were marked using this tag.

Notation: `[HNOISE]` (in upper-case letters)

Example: `sabena [HNOISE] four one report your heading`

A.1.7. Non-verbal Articulations

Definition: Non-verbal articulations such as confirmatory, surprise or hesitation sounds.

Notation: Limited set of expressions written in lower-case letters.

Example: `malaysian ah four is identified`

Word List: `ah hm ahm yeah aha nah ohh`[3]

[2] In some occasions less popular five-letter designators are also spelled out using ICAO phonetic spelling.

[3] In contrast to `ohh` as an expression of surprise, the notation `oh` is used for the meaning 'zero', as in `one oh one`.

A.1.8. Truncated Words

Definition: Words which are cut off either at the beginning or the end of the word due to stutters, full stops, or where the controller pressed the push-to-talk (PTT) button too late. This also applies to words that are interrupted by human noises (`[HNOISE]`). This notation is used when the word-part is understandable. Empty pauses within words are not marked.

Notation: The missing part of the word is replaced by an equals sign (=).

Examples: `good mor= good afternoon` (correction), `luf= lufthansa three two five` (stutter), `=bena four one` (PTT pressed too late), `sa= [HNOISE] =bena` (interruption by cough)

Exception: Words which are cut off either at the beginning or the end of the word due to fast speech or sloppy pronunciation are recorded according to standard spelling and not marked. If the word-part is too short to be identified, another notation is used (see below).

Exception Example: "goo'day" is transcribed as `good day`.

A.1.9. Word Fragments

Definition: Fragments of words that are too short so that no clear spelling of the fragment can be determined.

Notation: `[FRAGMENT]`

Example: `[FRAGMENT]`

A.1.10. Empty Utterances

Definition: Instances where the controller pressed the PTT button, said nothing at all, and released the button again.

Notation: `[EMPTY]`

Exception: If the utterance contains human noises produced by the speaker, the [HNOISE] tag is used.

A.1.11. Off-Talk

Definition: Speech that is neither addressed to the pilot nor part of the air traffic control communication.

Notation: Off-talk speech is transcribed and marked with opening and closing XML-style tags: `<OT> ... </OT>`

Example: `speedbird five nine zero <OT> ohh we are finished now </OT>`

A.1.12. Nonsensical Words

Definition: Clearly articulated word or word part that is not part of a dictionary and that also does not make any sense. This is usually a slip of the tongue and the speaker corrects the mistake.

Notation: `[NONSENSE]`

Example: `[NONSENSE] futura nine three three identified`

A.1.13. Foreign Language

Definition: Complete utterances, or parts thereof, given in a foreign language.

Notation: The foreign language part is not transcribed but is in its entirety replaced by adjacent XML-style tags: `<FL> </FL>`

Example: `<FL> </FL> break alitalia three seven zero report mach number`

Exception: Certain foreign language terms, such as greetings, are transcribed according to the spelling of that language, and are not tagged in any special way. A full list is given below.

Exception Examples `bonjour, tag, ciao`

Exception Word List: See Section A.2.4.

A.1.14. Unknown Words

Definition: Word or group of words that could not be understood or identified.

Notation: `[UNKNOWN]`

Example: `[UNKNOWN] five zero one bonjour cleared st prex`

A.2. Amendments to the Transcription Format

The actual language use in the recordings required the following additions to the above transcription format definitions.

A.2.1. Airline Telephony Designators

The following airline telephony designators cannot be found in the references cited above, but are nonetheless clearly identified.

A.2.1.1. Military Radio Call Signs

There was no special list for military aircraft call signs available. The following call signs were confirmed by an operational controller:

- `~i ~f ~o`
- `mission`

- nato
- spar
- steel

A.2.1.2. General Aviation Call Signs

In certain cases general aviation aircraft are addressed using the aircraft manufacturer and type number (e.g. fokker twenty eight). The following manufacturer names occurred:

- fokker
- ~b ~a (Short form for British Aerospace.)

A.2.1.3. Deviated Call Signs

In certain cases the controller uses a deviated or truncated version of the official call sign. The following uses occurred:

- bafair (Short form for belgian airforce.)
- netherlands (Short form for netherlands air force.)
- netherlands air (Short form for netherlands air force.)
- german air (Short form for german air force.)
- french air force (The official radio call sign is france air force.)
- israeli (Short form for israeli air force.)
- israeli air (Short form for israeli air force.)
- turkish (Short form for turkish airforce.)
- hapag (Short form for hapag lloyd.)
- french line (Short form for french lines.)
- british midland (This is the airline name. The radio call sign is midland.)
- berlin (Short form for air berlin.)
- algerie (Short form for air algerie.)
- hansa (Short form for lufthansa.)
- lufty (Short form for lufthansa.)
- luha (Short form for lufthansa.)
- france (Short form for airfrans.)
- meridiana (This is the airline name, which also used to be the radio call sign. The official call sign was changed to merair at some point in the past.)
- tunis air (This is the airline name. The radio call sign is tunair.)
- malta (Short form for air malta.)
- lauda (Short form for lauda air.)

A.2.1.4. Additional Verified Call Signs

The following call sign occurred and is also verified:

- london airtours (This call sign is listed only in the simulation manual [161].)

A.2.1.5. Additional Unverified Radio Call Signs

The following airline telephony designators could not be verified through any of the available resources. They are transcribed as understood by the transcriptionist on a best-guess basis and preceded by an at symbol (`@`).

`@aerovic`	`@cheeseburger`	`@indialook`	`@period`
`@alpha`	`@color`	`@ingishire`	`@roystar`
`@aviva`	`@devec`	`@jose`	`@sunwing`
`@bama`	`@foxy`	`@metavec`	`@taitian`
`@cheena`	`@hanseli`	`@nafamens`	`@tele`

A.2.2. Navigational Aids

A.2.2.1. Deviated Navaids

In certain cases the controller uses a deviated version of the official navaid name. The following uses occurred:

- `milano` (Local Italian version for `milan`.)
- `trasa` (Short form for `trasadingen`.)

A.2.2.2. Additional Verified Navaids

The following additional navaids occurred and are verified as they were occasionally spelled out by the controllers:

- `corna`
- `gotil`

A.2.3. Special Vocabulary and Abbreviations

The following ATC specific vocabulary and abbreviations occurred. This listing is most likely incomplete.

- `masp` (Minimum Aviation System Performance standards, pronounced as one word)
- `~r ~v ~s ~m` (Reduced Vertical Separation Minimum)
- `~c ~v ~s ~m` (Conventional Vertical Separation Minimum)
- `~i ~f runway` (Initial Fix runway)
- `sec` (sector)
- `freq` (frequency)

A.2.4. Foreign Language Greetings and Words

Due to their frequent occurrence the following foreign language greetings and words were transcribed, using a simplified spelling which avoids special characters:

- `hallo` (German for 'hello')
- `auf wiederhoren` (German for 'goodbye')
- `gruss gott` (German for 'hello')
- `servus` (German for 'hi')
- `guten morgen` (German for 'good morning')
- `guten tag` (German for 'hello')
- `adieu` (German for 'goodbye')
- `tschuss` (German for 'goodbye')
- `tschu` (German for 'goodbye')
- `danke` (German for 'thank you')
- `bonjour` (French for 'hello')
- `au revoir` (French for 'goodbye')
- `merci` (French for 'thank you')
- `hoi` (Dutch for 'hello')
- `dag` (Dutch for 'goodbye')
- `buongiorno` (Italian for 'hello')
- `arrivederci` (Italian for 'goodbye')
- `hejda` (Swedish for 'goodbye')
- `adios` (Spanish for 'goodbye')

Appendix B

Aeronautical Radio Channel Modeling and Simulation

In this appendix, the basic concepts in the modeling and simulation of the mobile radio channel are reviewed. The propagation channel is time-variant and dominated by multipath propagation, Doppler effect, path loss and additive noise. Stochastic reference models in the equivalent complex baseband facilitate a compact mathematical description of the channel's input-output relationship. The realization of these reference models as filtered Gaussian processes leads to practical implementations of frequency selective and frequency nonselective channel models. Three different small-scale area simulations of the aeronautical voice radio channel are presented and we demonstrate the practical implementation of a frequency flat fading channel. Based on a scenario in air/ground communication the parameters for readily available simulators are derived. The resulting outputs give insight into the characteristics of the channel and can serve as a basis for the design of digital transmission and measurement techniques. We conclude that the aeronautical voice radio channel is a frequency nonselective flat fading channel and that in most situations the frequency of the amplitude fading is band-limited to the maximum Doppler frequency.

Parts of this chapter have been published in K. Hofbauer and G. Kubin, "Aeronautical voice radio channel modelling and simulation—a tutorial review," in *Proceedings of the International Conference on Research in Air Transportation (ICRAT)*, Belgrade, Serbia, Jul. 2006.

B.1. Introduction

Radio channel modeling has a long history and is still a very active area of research. This is especially the case with respect to terrestrial mobile radio communications and wideband data communications due to commercial interest. However, the results are not always directly transferable to the aeronautical domain. A comprehensive up-to-date literature review on channel modeling and simulation with the aeronautical radio in mind is provided in [134]. It is highly recommended as a pointer for further reading and its content is not repeated herein. In this appendix, we review the general concepts of radio channel modeling and demonstrate the application of three readily available simulators to the aeronautical voice channel.

B.2. Basic Concepts

This and the following section are based on the work of Pätzold [132] and provide a summary of the basic characteristics, the modeling, and the simulation of the mobile radio channel. Another comprehensive treatment on this extensive topic is given in [133].

B.2.1. Amplitude Modulation and Complex Baseband

The aeronautical voice radio is based on the double-sideband amplitude modulation (DSB-AM, A3E or simply AM) of a sinusoidal, unsuppressed carrier [65]. An analog baseband voice signal $x(t)$ which is band-limited to a bandwidth f_m modulates the amplitude of a sinusoidal carrier with amplitude A_0, carrier frequency f_c and initial phase φ_0. The modulated high frequency (HF) signal $x_{AM}(t)$ is defined as

$$x_{AM}(t) = (A_0 + kx(t))\cos(2\pi f_c t + \varphi_0)$$

with the modulation depth

$$m = \frac{|kx(t)|_{max}}{A_0} \leq 1$$

The real-valued HF signal can be equivalently written using complex notation and $\omega_c = 2\pi f_c$ as

$$x_{AM}(t) = Re\left\{(A_0 + kx(t))e^{j\omega_c t}e^{j\varphi_0}\right\} \tag{B.1}$$

Under the assumption that $f_c \gg f_m$ the HF signal can be demodulated and the input signal $x(t)$ reconstructed by detecting the envelope of the modulated sine wave. The absolute value is low-pass filtered and the original amplitude of the carrier is subtracted.

$$x(t) = \frac{1}{k}([|x_{AM}(t)|]_{LP} - A_0)$$

Figure B.1 shows the spectra of the baseband signal and the corresponding HF signal. Since the baseband signal is, by definition, low-pass filtered, the HF signal is a

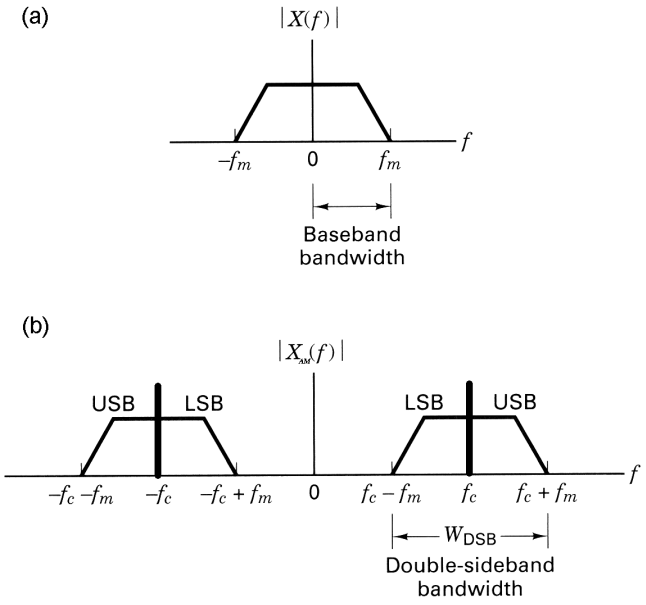


Figure B.1.: Signal spectra of (a) the baseband signal $x(t)$ and (b) the HF signal $x_{AM}(t)$ with a carrier at f_c and symmetric upper and lower sidebands (Source: [139], modified).

bandpass signal and contains energy only around the carrier frequency and the lower and upper sidebands LSB and USB.

In general, any real bandpass signal $s(t)$ can be represented as the real part of a modulated complex signal,

$$s(t) = Re\left\{g(t)e^{j\omega_c t}\right\} \tag{B.2}$$

where $g(t)$ is called the equivalent complex baseband or complex envelope of $s(t)$ [139]. The complex envelope $g(t)$ is obtained by downconversion of the real passband signal $s(t)$, namely

$$g(t) = (s(t) + j\hat{s}(t))e^{-j\omega_C t}$$

with $\hat{s}(t)$ being the Hilbert transform of $s(t)$. The Hilbert transform removes the negative frequency component of $s(t)$ before downconversion [86]. A comparison of (B.1) and (B.2) reveals that the complex envelope $g_{AM}(t)$ of the amplitude modulated HF signal $x_{AM}(t)$ simplifies to

$$g_{AM}(t) = (A_0 + kx(t))e^{j\varphi_0}$$

The signal $x(t)$ is reconstructed from the equivalent complex baseband of the HF signal by demodulation with

$$x(t) = \frac{1}{k}\left(|g_{AM}(t)| - A_0\right) \tag{B.3}$$

The complex baseband signal can be pictured as a time-varying phasor or vector in a rotating complex plane. The rotating plane can be seen as a coordinate system for the vector, which rotates with the angular velocity ω_c.

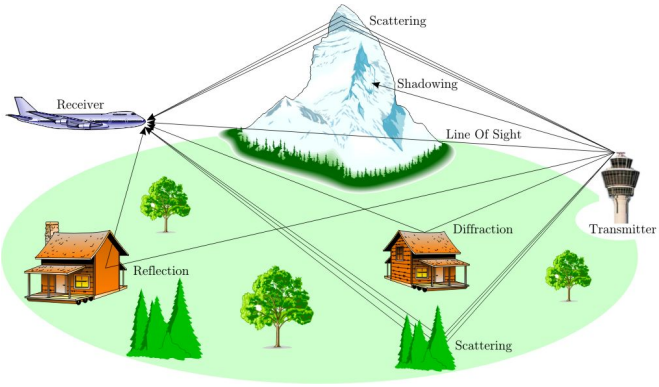


Figure B.2.: Multipath propagation in an aeronautical radio scenario (Source: [162]).

In order to represent the HF signal as a discrete-time signal, it must be sampled with a frequency of more than twice the carrier frequency. This leads to a large number of samples and thus makes numerical simulation difficult even for very short signal durations. Together with the carrier frequency ω_c the complex envelope $g_{AM}(t)$ fully describes the HF signal $x_{AM}(t)$. The complex envelope $g_{AM}(t)$ has the same bandwidth $[-f_m; f_m]$ as the baseband signal $x(t)$. As a consequence it can be sampled with a much lower sampling frequency, which facilitates efficient numerical simulation without loss of generality. Most of the existing channel simulations are based on the complex baseband signal representation.

B.2.2. Mobile Radio Propagation Channel

Proakis [65] defines the communication channel as "... the physical medium that is used to send the signal from the transmitter to the receiver." Radio channel modeling usually also includes the transmitting and receiving antennas in the channel model.

B.2.2.1. Multipath Propagation

The transmitting medium in radio communications is the atmosphere or free space, into which the signal is coupled as electromagnetic energy by an antenna. The received electromagnetic signal can be a superposition of a line-of-sight path signal and multiple waves coming from different directions. This effect is known as multipath propagation. Depending on the geometric dimensions and the properties of the objects in a scene, an electromagnetic wave can be reflected, scattered, diffracted or absorbed on its way to the receiver.

From hereon we assume, without loss of generality, that the ground station transmits and the aircraft receives the radio signal. The effects treated in this paper are identical for both directions. As illustrated in Figure B.2, reflected waves have to travel a longer distance to the aircraft and therefore arrive with a time-delay compared to the line-of-sight signal. The received signal is spread in time and the channel is said to be time dispersive. The time delays correspond to phase shifts in between the superimposed

waves and lead to constructive or destructive interference depending on the position of the aircraft. As both the position and the phase shifts change constantly due to the movement of the aircraft, the signal undergoes strong pseudo-random amplitude fluctuations and the channel becomes a fading channel.

The multipath spread T_m is the time delay between the arrival of the line-of-sight component and the arrival of the latest scattered component. Its inverse $B_{CB} = \frac{1}{T_M}$ is the coherence bandwidth of the channel. If the frequency bandwidth W of the transmitted signal is larger than the coherence bandwidth ($W > B_{CB}$), the channel is said to be frequency selective. Otherwise, if $W < B_{CB}$, the channel is frequency nonselective or flat fading. This means that all the frequency components of the received signal are affected by the channel in the same way [65].

B.2.2.2. Doppler Effect

The so-called Doppler effect shifts the frequency content of the received signal due to the movement of the aircraft relative to the transmitter. The Doppler frequency f_D, which is the difference between the transmitted and the received frequency, is dependent on the angle of arrival α of the electromagnetic wave relative to the heading of the aircraft.

$$f_D = f_{D,max} \cos(\alpha)$$

The maximum Doppler frequency $f_{D,max}$, which is the largest possible Doppler shift, is given by

$$f_{D,max} = \frac{v}{c} f_c \tag{B.4}$$

where v is the aircraft speed, f_c the carrier frequency and $c = 3 \cdot 10^8 \frac{\mathrm{m}}{\mathrm{s}}$ the speed of light.

The reflected waves arrive not only with different time-delays compared to the line-of-sight signal, but as well from different directions relative to the aircraft heading (Figure B.2). As a consequence, they undergo different Doppler shifts. This results in a continuous distribution of frequencies in the spectrum of the signal and leads to the so-called Doppler power spectral density or simply Doppler spectrum.

B.2.2.3. Channel Attenuation

The signal undergoes significant attenuation during transmission. The path loss is dependent on the distance d and the obstacles between transmitter and receiver. It is proportional to $\frac{1}{d^p}$, with the pathloss exponent p in the range of $2 \leq p < 4$. In the optimal case of line-of-sight free space propagation $p = 2$.

B.2.2.4. Additive Noise

During transmission additive noise is imposed onto the signal. The noise results, among others, from thermal noise in electronic components, from atmospheric noise or radio channel interference, or from man-made noise such as engine ignition noise.

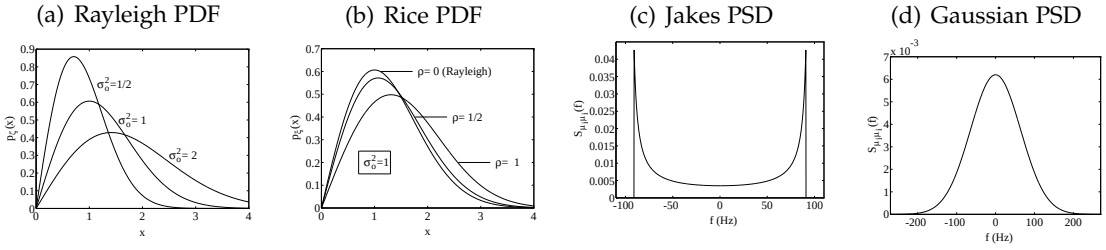


Figure B.3.: Probability density functions (PDF) and power spectral densities (PSD, $f_{D,max} = 91\,\text{Hz}$, $\sigma_0^2 = 1$) for Rayleigh and Rice channels (Source: [132]).

B.2.2.5. Time Dependency

Most of the parameters described in this section vary over time due to the movement of the aircraft. As a consequence the response of the channel to a transmitted signal also varies, and the channel is said to be time-variant.

B.2.3. Stochastic Terms and Definitions

The following section recapitulates some basic stochastic terms in order to clarify the nomenclature and notation used herein. The reader is encouraged to refer to [132] for exact definitions.

Let the event A be a collection of a number of possible outcomes s of a random experiment, with the real number $P(A)$ being its probability measure. A random variable μ is a mapping that assigns a real number $\mu(s)$ to every outcome s. The cumulative distribution function

$$F_\mu(x) = P(\mu \leq x) = P(\{s|\mu(s) \leq x\})$$

is the probability that the random variable μ is less or equal to x. The probability density function (PDF, or simply density) $p_\mu(x)$ is the derivative of the cumulative distribution function,

$$p_\mu(x) = \frac{dF_\mu(x)}{dx}$$

The most common probability density functions are the uniform distribution, where the density is constant over a certain interval and is zero outside, and the Gaussian distribution or normal distribution $N(m_\mu, \sigma_\mu^2)$, which is determined by the two parameters expected value m_μ and variance σ_μ^2.

With μ_1 and μ_2 being two statistically independent normally distributed random variables with identical variance σ_0^2, the new random variable $\zeta = \sqrt{\mu_1^2 + \mu_2^2}$ represents a Rayleigh distributed random variable (Figure B.3(a)). Given an additional real parameter ρ, the new random variable $\xi = \sqrt{(\mu_1 + \rho)^2 + \mu_2^2}$ is Rice or Rician distributed (Figure B.3(b)). A random variable $\lambda = e^\mu$ is said to be lognormally distributed. A multiplication of a Rayleigh and a lognormally distributed random variable $\eta = \zeta\lambda$ leads to the so-called Suzuki distribution.

A stochastic process $\mu(t,s)$ is a collection of random variables, which is indexed by a time index t. At a fixed time instant $t = t_0$, the value of a random process, $\mu(t_0,s)$, is a random variable. On the other hand, in the case of a fixed outcome $s = s_0$ of a random experiment, the value of the stochastic process $\mu(t,s_0)$ is a time function, or signal, that corresponds to the outcome s_0. As is common practice, the variable s is dropped in the notation for a stochastic process and $\mu(t)$ written instead. With $\mu_1(t)$ and $\mu_2(t)$ being two real-valued stochastic processes, a complex-valued stochastic process is defined by $\mu(t) = \mu_1(t) + j\mu_2(t)$. A stochastic process is called stationary if its statistical properties are invariant to a shift in time. The Fourier transform of the autocorrelation function of such a stationary process defines the power spectral density or power density spectrum of the stochastic process.

B.3. Radio Channel Modeling

Section B.2.2.1 illustrated multipath propagation from a geometrical point of view. However, geometrical modeling of the multipath propagation is possible only to a very limited extent. It requires detailed knowledge of the geometry of all objects in the scene and their electromagnetic properties. The resulting simulations are time consuming to set up and computationally expensive, and a number of simplifications have to be made. Furthermore the results are valid for the specific situation only and cannot always be generalized. As a consequence, a stochastic description of the channel and its properties is widely used. It focuses on the distribution of parameters over time instead of trying to predict single values. This class of stochastic channel models is the subject of the following investigations.

In large-scale areas with dimensions larger than tens of wavelengths of the carrier frequency f_c, the local mean of the signal envelope fluctuates mainly due to shadowing and is found to be approximately lognormally distributed. This slow fading is important for channel availability, handover, and mobile radio network planning.

More important for the design of a digital transmission technique is the fast signal fluctuation, the fast fading, which occurs within small areas. As a consequence, we focus on models that are valid for small-scale areas, where we can assume the path loss and the local mean of the signal envelope due to shading, etc., to be constant. Furthermore we assume for the moment a frequency nonselective channel and, for mathematical simplicity, the transmission of an unmodulated carrier. A more extensive treatment can be found in [132], on which this section is based on.

B.3.1. Stochastic Mutlipath Reference Models

The sum $\mu(t)$ of all scattered components of the received signals can be assumed to be normally distributed. If we let $\mu_1(t)$ and $\mu_2(t)$ be zero-mean statistically independent Gaussian processes with variance σ_0^2, then the sum of the scattered components is given in complex baseband representation as a zero-mean complex Gaussian process $\mu(t)$ and is defined by

$$\text{Scatter:} \qquad \mu(t) = \mu_1(t) + j\mu_2(t) \tag{B.5}$$

The line-of-sight (LOS) signal component $m(t)$ is given by

$$\text{LOS:} \qquad m(t) = A_0 e^{j(2\pi f_D + \varphi_0)}, \tag{B.6}$$

again in complex baseband representation. The superposition $\mu_m(t)$ of both signals is

$$\text{LOS+Scatter:} \qquad \mu_m(t) = m(t) + \mu(t) \tag{B.7}$$

Depending on the surroundings of the transmitter and the receiver, the received signal consists of either the scatter components only or a superposition of LOS and scatter components. In the first case (i.e. (B.5)) the magnitude of the complex baseband signal $|\mu(t)|$ is Rayleigh distributed. Its phase $\angle(\mu(t))$ is uniformly distributed over the interval $[-\pi; \pi)$. This type of a Rayleigh fading channel is predominant in regions where the LOS component is blocked by obstacles, such as in urban areas with high buildings, etc.

In the second case where a LOS component and scatter components are present (i.e. (B.7)), the magnitude of the complex baseband signal $|\mu(t) + m(t)|$ is Rice distributed. The Rice factor k is determined by the ratio of the power of the LOS and the scatter components, where $k = \frac{A_0^2}{2\sigma_0^2}$. This Rice fading channel dominates the aeronautical radio channel.

One can derive the probability density of amplitude and phase of the received signal based on the Rice or Rayleigh distributions. As a further step, it is possible to compute the level crossing rate and the average duration of fades, which are important measures required for the optimization of coding systems in order to address burst errors. The exact formulas can be found in [132] and are not reproduced herein.

The power spectral density of the complex Gaussian random process in (B.7) corresponds to the Doppler power spectral density when considering the power of all components, their angle of arrival and the directivity of the receiving antenna. Assuming a Rayleigh channel with no LOS component, propagation in a two-dimensional plane and uniformly distributed angles of arrival, one obtains the so-called Jakes power spectral density as the resulting Doppler spectrum. Its shape is shown in Figure B.3(c).

However, both theoretical investigations and measurements have shown that the assumption that the angle of arrival of the scattered components is uniformly distributed does in practice not hold for aeronautical channels. This results in a Doppler spectrum which is significantly different from the Jakes spectrum [163]. The Doppler power spectral density is therefore better approximated by a Gaussian power spectral density, which is plotted in Figure B.3(d). For nonuniformly distributed angles of arrival, as with explicit directional echos, the Gaussian Doppler PSD is unsymmetrical and shifted away from the origin. The characteristic parameters describing this spectrum are the average Doppler shift (the statistic mean) and the Doppler spread (the square root of the second central moment) of the Doppler PSD .

B.3.2. Realization of the Reference Models

The above reference models are based on colored Gaussian random processes. The realization of these processes is not trivial and leads to the theory of deterministic

processes. Mostly two fundamental methods are applied in the literature in order to generate colored Gaussian processes. In the filter method, white Gaussian noise is filtered by an ideal linear time-invariant filter with the desired power spectrum. In the Rice method an infinite number of weighted sinusoids with equidistant frequencies and random phases are superimposed. In practice both methods can only approximate the colored Gaussian process. Neither an ideal filter nor an infinite number of sinusoids can be realized. A large number of algorithms used to determine the actual parameters of the sinusoids in the Rice method exist. The methods approximate the Gaussian processes with a sum of a limited number of sinusoids, thus considering the computational expense [132]. For the filter method on the other hand, the problem boils down to filter design with its well-understood limitations.

B.3.3. Frequency Nonselective Channel Models

In frequency nonselective flat fading channels, all frequency components of the received signal are affected by the channel in the same way. The channel is modeled by a multiplication of the transmitted signal with a suitable stochastic model process. The Rice and Rayleigh processes described in Section B.3.1 can serve as statistical model processes.

However it has been shown that the Rice and Rayleigh processes often do not provide enough flexibility to adapt to the statistics of real world channels. This has led to the development of more versatile stochastic model processes such as the Suzuki process and its variations (a product of a lognormal distribution for the slow fading and a Rayleigh distribution for the fast fading), the Loo Model with its variations, and the generalized Rice process.

B.3.4. Frequency Selective Channel Models

Where channel bandwidth and data rate increase, the propagation delays can no longer be ignored as compared to the symbol interval. The channel is then said to be frequency selective and (over time) the different frequency components of a signal are affected differently by the channel.

B.3.4.1. Tapped Delay Line Structure

For the modeling of a frequency selective channel, a tapped delay line structure is typically applied as reference model (Figure B.4). The ellipses model of Parsons and Bajwa [133] shows that all reflections and scatterings from objects located on an ellipse, with the transmitter and receiver in the focal points, undergo the same time delay. This leads to a complex Gaussian distribution of the received signal components for a given time delay, assuming a large number of objects with different reflection properties in the scene and applying the central limit theorem. As a consequence, the tap weights $c_i(t)$ of the single paths are assumed to be uncorrelated complex Gaussian processes. It is shown in Section B.2.3 that the amplitudes of the complex tap weights are then

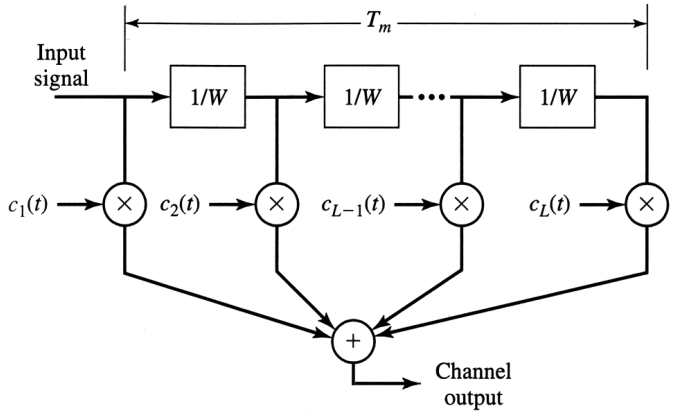


Figure B.4.: Tapped delay line structure as a frequency selective and time variant channel model (Source: [132]). W is the bandwidth of the transmitted signal, $c_i(t)$ are uncorrelated complex Gaussian processes.

either Rayleigh or Rice distributed, depending on the mean of the Gaussian processes. An analytic expression for the phases of the tap weights can be found in [132].

B.3.4.2. Linear Time-Variant System Description

The radio channel can be modeled as a linear time-variant system, with input and output signals in the complex baseband representation. The system can be fully described by its time-variant impulse response $h(t, \tau)$. In order to establish a statistical description of the input/output relation of the above system, the channel is further considered as a stochastic system, with $h(t, \tau)$ as its stochastic system function.

These input/output relations of the stochastic channel can be significantly simplified assuming that the impulse response $h(t, \tau)$ is wide sense stationary[1] (WSS) in its temporal dependence on t, and assuming that scattering components with different propagation delays τ are statistically uncorrelated (uncorrelated scattering (US)). Based on these two assumptions, Bello proposed in 1963 the class of WSSUS models. They are nowadays widely used and are of great importance in channel modeling. They are based on the tapped delay line structure and allow the computation of all correlation functions, power spectral densities and properties such as Doppler and delay spread, etc., from a given scattering function. The scattering function may be obtained by the measurement of real channels, by specification, or both. For example, the European working group 'COST 207' published scattering functions in terms of delay power spectral densities and Doppler power spectral densities for four propagation environments which are claimed to be typical for mobile cellular communication.

[1] Measurements have shown that this assumption is valid for areas smaller than tens of wavelengths of the carrier frequency f_c.

B.3.5. AWGN Channel Model

The noise that is added to the transmitted signal during transmission is typically represented as an additive white Gaussian noise (AWGN) process. The main parameter of the model is the variance σ_0^2 of the Gaussian process, which together with the signal power defines the signal-to-noise ratio (SNR) of the output signal [65]. The AWGN channel is usually included as an additional block after the channel models described earlier.

B.4. Aeronautical Voice Channel Simulation

This section aims to present three different simulators which implement the above radio channel models. As mentioned earlier, the models are based on a small-scale area assumption where path loss and shadowing are assumed to be constant. We first define a simulation scenario based on which we show the simulators' input parameters and the resulting channel output. We use as example the aeronautical VHF voice radio channel between a fixed ground station and a general aviation aircraft which is flying at its maximum speed.

The input and output signals of all three simulators are equivalent complex baseband signals, and the same Matlab-based pre- and post-processing of the signals is used for all simulators. The processing consists of bandpass pre- and post-filtering, conversion to and from complex baseband, and amplitude modulation and demodulation.

B.4.1. Simulation Scenario and Parameters

For air-ground voice communication in civil air traffic control, the carrier frequency f_c is within a range from 118 MHz to 137 MHz, the 'very high frequency' (VHF) band. The 760 channels are spaced 25 kHz apart. The channel spacing is reduced to 8.33 kHz in specific regions of Europe in order to increase the number of available channels to a theoretical maximum of 2280. According to specification, the frequency response of the transmitter is required to be flat between 0.3 kHz to 2.5 kHz with a sharp cut-off below and above this frequency range [97], resulting in an analog channel bandwidth $W = 2.2\,\text{kHz}$.

For the simulations, we assume a carrier with amplitude $A_0 = 1$, frequency $f_c = 120\,\text{MHz}$ and initial phase $\varphi_0 = \frac{\pi}{4}$, a channel spacing of 8.33 kHz, a modulation depth $m = 0.8$ and an input signal which is band-limited to $f_l = 300\,\text{Hz}$ to $f_m = 2.5\,\text{kHz}$. For the illustrations we use a purely sinusoidal baseband input signal $x(t) = \sin(2\pi f_a t)$ with $f_a = 500\,\text{Hz}$, which is sampled with a frequency of $f_{sa} = 8000\,\text{Hz}$ and bandpass filtered according to the above specification. Figure B.5 shows all values that the amplitude modulated signal $x_{AM}(t)$ takes on during the observation interval in the equivalent complex baseband representation $g_{AM}(t)$. The white circle represents the unmodulated carrier signal, which is a single point in the equivalent complex baseband. A short segment of the magnitude of the signal, $|g_{AM}(t)|$, is also shown in the figure.

In the propagation model a general aviation aircraft with a speed of $v = 60\,\frac{\text{m}}{\text{s}}$ is assumed. Using (B.4), this results in a maximum Doppler frequency of $f_{D,max} = 24\,\text{Hz}$.

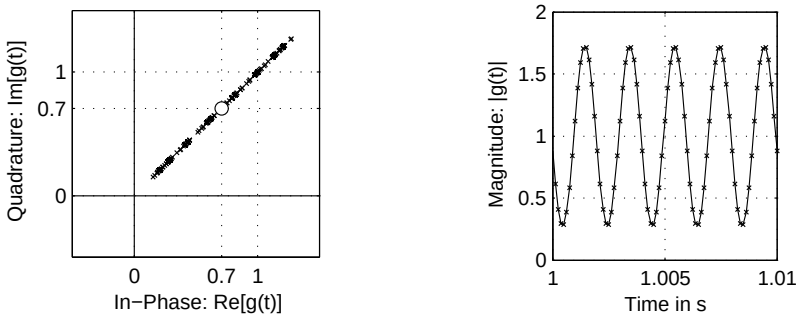


Figure B.5.: Sinusoidal AM signal in equivalent complex baseband. Left: In-phase and quadrature components. The white circle indicates an unmodulated carrier. Right: The magnitude of $g_{AM}(t)$.

Given the carrier frequency f_c, the wavelength is $\lambda = \frac{c}{f_c} = 2.5\,\text{m}$. This distance λ is covered in $t_\lambda = 0.0417\,\text{s}$. Furthermore, we assume that the aircraft flies at a height of $h_2 = 3000\,\text{m}$ and at a distance of $d = 10\,\text{km}$ from the ground station. The ground antenna is considered to be mounted at a height of $h_1 = 20\,\text{m}$. The geometric path length difference Δl between the line-of-sight path and the dominant reflection along the vertical plane on the horizontal flat ground evaluates to

$$\Delta l = \sqrt{h_1^2 + \left(\frac{h_1 d}{h_1 + h_2}\right)^2} + \sqrt{h_2^2 + \left(d - \frac{h_1 d}{h_1 + h_2}\right)^2} - \sqrt{d^2 + (h_2 - h_1)^2} = 11.5\,\text{m}$$

which corresponds to a path delay of $\Delta\tau = 38.3\,\text{ns}$. In a worst case scenario with a multipath spread of $T_m = 10\Delta\tau$, the coherence bandwidth is still $B_{CB} = 2.6\,\text{MHz}$. With $B_{CB} \gg W$, according to Section B.2.2.1 the channel is surely frequency nonselective. Worst-case multipath spreads of $T_m = 200\,\mu\text{s}$ as reported in [163] cannot be explained with a reflection in the vertical plane, but only with a reflection on far-away steep slopes. In these rare cases, the resulting coherence bandwidth is in the same order of magnitude as the channel bandwidth.

We cannot confirm the rule of thumb given in [163] where $\Delta l \approx h_2$ given $d \gg h_2$. For example, a typical case for commercial aviation where $h_1 = 30\,\text{m}, h_2 = 10000\,\text{m}$ and $d = 100\,\text{km}$ results in a path difference of $\Delta l = 6.0\,\text{m}$. In the special case of a non-elevated ground antenna with $h_1 \approx 0$ the path delay vanishes. In contrast, large path delays only occur in situations with large h_1 and h_2 and small d such as in air-to-air communication, which is not considered herein.

The Rician factor k is assumed to be $k = 12\,\text{dB}$, which corresponds to a fairly strong line-of-sight signal [163].

B.4.2. Mathworks Communications Toolbox Implementation

The Mathworks Communications Toolbox for Matlab [164] implements a multipath fading channel model. The simulator supports multiple fading paths, of which the first is Rice or Rayleigh distributed and the subsequent paths are Rayleigh distributed. The

Doppler spectrum is approximated by the Jakes spectrum. As shown in Section B.3.1, the Jakes Doppler spectrum is not suitable for the aeronautical channel. The preferable Gaussian Doppler spectrum is unfortunately not supported by the simulator. The toolbox provides a convenient tool for the visualization of the impulse and frequency response, the gain and phasor of the multipath components, and the evolution of these quantities over time.

In terms of implementation, the toolbox models the channel as a time-variant linear FIR filter. Its tap-weights $g(m)$ are given by a sampled and truncated sum of shifted sinc functions. They are shifted by the path delays τ_k of the k^{th} path, weighted by the average power gain p_k of the corresponding path, and weighted by a random process $h_k(n)$. The uncorrelated random processes $h_k(n)$ are filtered Gaussian random processes with a Jakes power spectral density. This results in

$$g(m) = \sum_k \operatorname{sinc}\left(\frac{\tau_k}{1/f_{sa}} - m\right) h_k(n) p_k.$$

The equation shows once again that when all path delays are small as compared to the sample period, the sinc terms coincide. This results in a filter with only one tap and consequently in a frequency nonselective channel.

In our scenario the channel is frequency-flat, and a model according to Section B.3.3 with one Rician path is appropriate. The only necessary input parameters for the channel model are f_{sa}, $f_{D,max}$ and k.

The output of the channel for the sinusoidal input signal as defined above is shown in Figure B.6(a). The demodulated signal (using (B.3) and shown in Figure B.6(b)) reveals the amplitude fading of the channel due to the Rician distribution of the signal amplitude. It is worthwhile noticing that the distance between two maxima is roughly one wavelength λ. This fast fading results from the superposition of the line-of-sight component and the multitude of scattered components with Gaussian distribution. As shown in Figure B.6(c), bandpass filtering the received signal according to the specified channel bandwidth does not remove the amplitude fading.

The toolbox also allows a model structure with several discrete paths similar to Figure B.4. One can specify the delay and the average power of each path. A scenario similar to the first one with two distinct paths is shown for comparison. We define one Rician path with a Rician factor of $k = 200$. This means that it contains only the line of sight signal and no scattering. We furthermore define one Rayleigh path with a relative power of -12 dB and a time delay of $\Delta\tau = 38.3$ ns, both relative to the LOS path.

Due to the small bandwidth of our channel, the results are equivalent to the first scenario. Figure B.7 shows the time-variation of the power of the two components, with the total power being normalized to 0 dB.

B.4.3. The Generic Channel Simulator Implementation

The Generic Channel Simulator (GCS) is a radio channel simulator which was developed between 1994 and 1998 under contract of the American Federal Aviation Administration (FAA). Its source code and documentation is provided in [165]. The

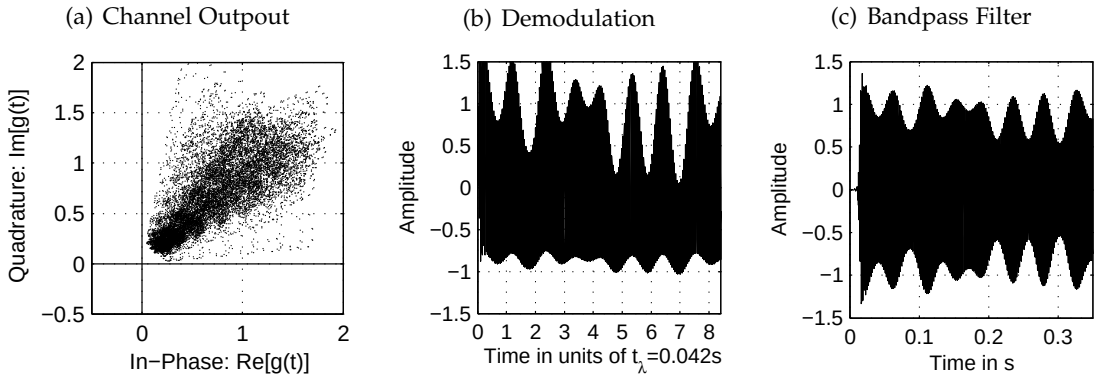


Figure B.6.: Received signal at different processing stages. Received signal (channel output of Mathworks Communication Toolbox and an observation interval of 2 s) (a) in equivalent complex baseband, (b) after demodulation, and (c) after bandpass filtering.

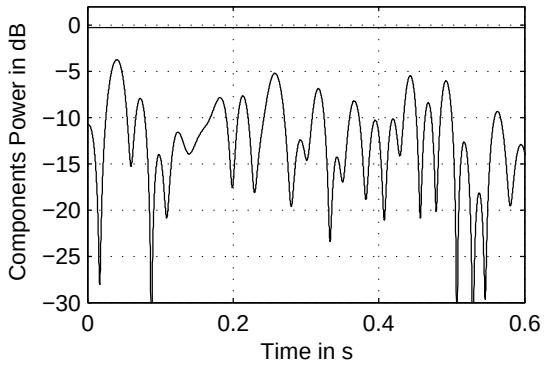


Figure B.7.: Power of the line-of-sight component (top) and the Rayleigh distributed scattered components (bottom).

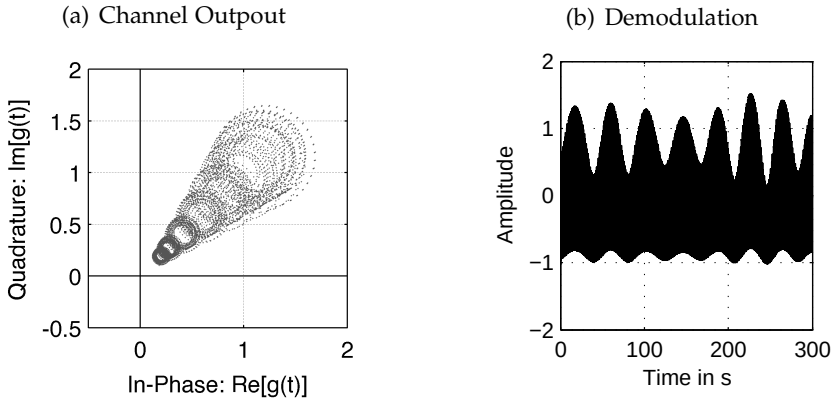


Figure B.8.: Generic Channel Simulator: Received signal in an observation interval of 300 s (channel output) (a) in equivalent complex baseband and (b) after demodulation.

GCS written in ANSI C and provides a graphical MOTIF-based interface and a command line interface to enter the model parameters. Data input and output files are in a binary floating point format and contain the signal in equivalent complex baseband representation. The last publicly available version of the software dates back to 1998. This version requires a fairly complex installation procedure and a number of adjustments in order to enable compiling of the source code on current operating systems. We provide some advice on how to install the software on the Mac OS X operating system and how to interface the simulator with Matlab [166].

The GCS allows the simulation of various types of mobile radio channels, the VHF air/ground channel among others. Similar to the Mathworks toolbox, the GCS simulates the radio channel by a time-variant IIR filter. The scatter path delay power spectrum shape is approximated by a decaying exponential multiplying a zeroth-order modified Bessel function, the Doppler power spectrum is assumed to have a Gaussian shape.

In the following example we use a similar setup as in the second scenario in Section B.4.2, a discrete line-of-sight signal and a randomly distributed scatter path with a power of -12 dB. With the same geometric configuration as above, the GCS software confirms our computed time delay of $\Delta\tau = 38.3\,\text{ns}$.

Using the same parameters for speed, geometry, frequency, etc., as in the scenarios described above, we obtain a channel output as shown in Figure B.8.

The time axis in the plot of the demodulated signal is very different as compared to the one in Figure B.6(c). The amplitude fading of the signal has a similar shape as before but is in the order of three magnitudes slower than observed in Section B.4.2. This contradicting result cannot be explained by the differing model assumptions, nor does it coincide with first informal channel measurements that we pursued.

We believe that the out-dated source code of the GCS has legacy issues which lead to problems with current operating system and compiler versions.

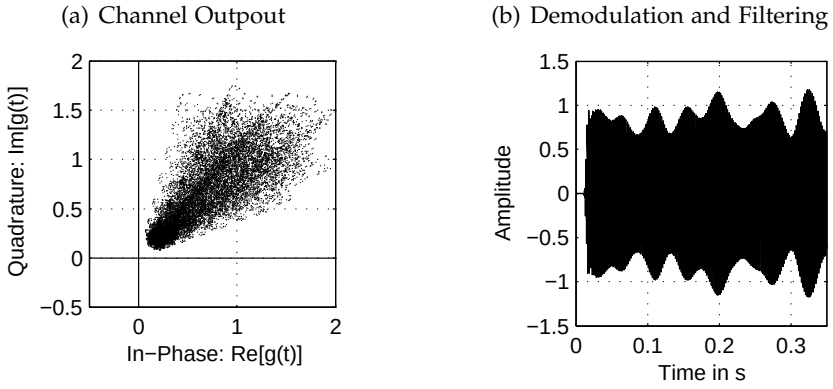


Figure B.9.: Reference channel model with Jakes PSD: Received signal (channel output) (a) in equivalent complex baseband and (b) after demodulation and bandpass filtering.

B.4.4. Direct Reference Model Implementation

The third channel simulation is based on a direct implementation of the concept described in Section B.3.3 with the Matlab routines provided in [132]. We model the channel by multiplying the input signal with the complex random process $\mu_m(t)$ as given in (B.7). The random processes are generated with the sum of sinusoids approach as discussed in Section B.3.2. Figure B.9 shows the channel output using a Jakes Doppler PSD and a Rician reference model. The result is very similar to the one obtained with the Mathworks toolbox.

In a second example, a Gaussian distribution is used for the Doppler power spectral density instead of the Jakes model. The additional parameter $f_{D,co}$ describes the width of the Gaussian PSD by its 3 dB cut-off frequency. The value is arbitrarily set to $f_{D,co} = 0.3 f_{D,max}$. This corresponds to a fairly small Doppler spread of $B = 6.11$ Hz, compared to the Jakes PSD with $B = 17$ Hz. Figure B.10 shows the resulting discrete Gaussian PSD. The channel output in Figure B.11 confirms the smaller Doppler spread by a narrower lobe in the scatter plot. The amplitude fading is by a factor of two slower than with the Jakes PSD.

B.5. Discussion

We further examine the path gain of the frequency nonselective flat fading channel simulated in Section B.4.2. This path gain is plotted in Figure B.12(a) and corresponds to the amplitude fading of the received signal shown in Figure B.6(b). Considering the time-variant path gain as a time series, its magnitude spectrum (shown in Figure B.12(b)) reveals that the maximum frequency with which the path gain changes approximately coincides with the maximum Doppler frequency $f_{D,max} = \frac{v}{c} f_c$ of the channel. The amplitude fading is band-limited to the maximum Doppler frequency.

This fact can be explained with the concept of beat as known in acoustics [167].

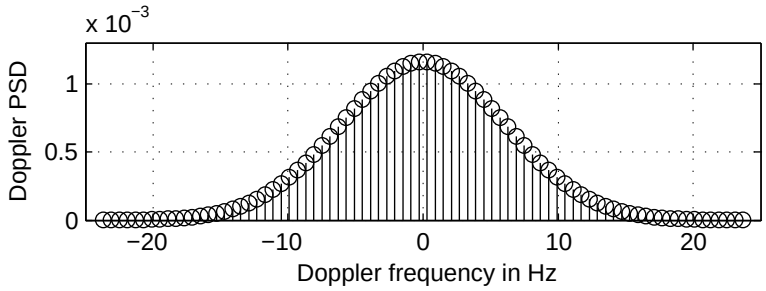


Figure B.10.: Discrete Gaussian Doppler PSD with $f_{D,max} = 24\,\text{Hz}$ and a 3-dB-cut-off frequency of $f_{D,co} = 7.2\,\text{Hz}$.

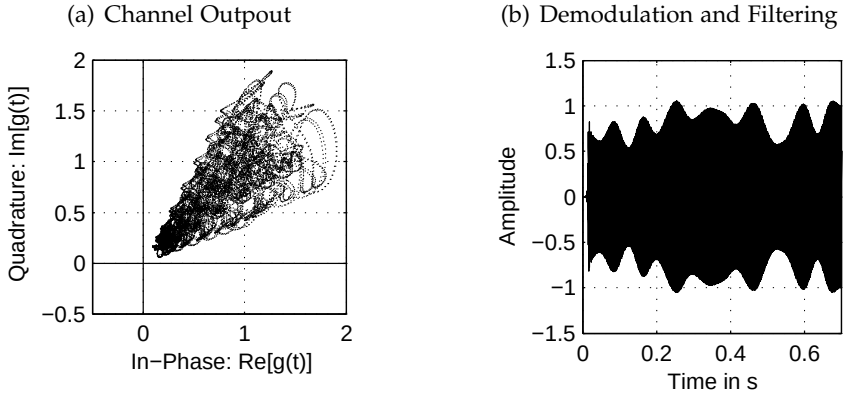


Figure B.11.: Reference channel model with Gaussian PSD: Received signal (channel output) (a) in equivalent complex baseband and (b) after demodulation and bandpass filtering.

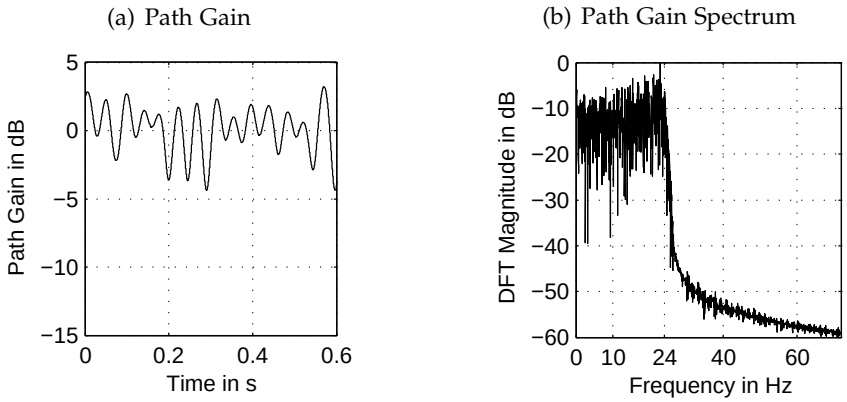


Figure B.12.: Time-variant path gain and its magnitude spectrum for a flat fading channel with $f_{D,max} = 24\,\text{Hz}$.

The superposition of two sinusoidal waves with slightly different frequencies f_1 and f_2 leads to a special type of interference, where the envelope of the resulting wave modulates with a frequency of $f_1 - f_2$. The maximum frequency difference $f_1 - f_2$ between a scatter component and the carrier frequency f_c with which we demodulate in the simulation is given by the maximum Doppler frequency. This explains the band-limitation of the amplitude fading.

In a real world system a potentially coherent receiver may demodulate the HF signal with a reconstructed carrier frequency $\hat{f}_c$ which is already Doppler shifted. In this case, the maximum frequency difference between Doppler shifted carrier and Doppler shifted scatter component is $2f_{D,max}$. This maximum occurs when the aircraft points towards the ground station so that the LOS signal arrives from in front of the aircraft, and when at the same time a scatter component arrives from the back of the aircraft [163]. We can thereof conclude that the amplitude fading of the frequency nonselective aeronautical radio channel is band-limited to twice the maximum Doppler frequency $f_{D,max}$.

For a channel measurement system as presented in Chapter 8 the maximum frequency of the amplitude fading entails that the fading of the channel has to be sampled with at least double the frequency, that means with a sampling frequency of $f_{ms} \geq 4f_{D,max}$, to avoid aliasing. With the parameters for a general aviation aircraft used throughout this appendix, this means that the amplitude scaling has to be sampled with a frequency of $f_{ms} > 96\,\text{Hz}$ or, in terms of an audio sample rate $f_{sa} = 8\,\text{kHz}$, at least every 83 samples. For a measurement system based on maximum length sequences (MLS, see Chapter 8) this means that the MLS length should be no longer than $L = 2^n - 1 = 63$ samples.

A factor that is not considered in the presented channel models is the likely presence of automatic gain controls (AGC) in the radio transmitter and receiver. Since the presented models are propagation channel models, transceiver components are by definition not part of these models. Transceiver modeling is in fact a discipline by itself and is outside the scope of this work. However, it should be noted that an AGC in the receiver might partially compensate for the amplitude modulation induced by the fading channel. Figure B.13(a) and B.13(b) show the demodulated signal of Section B.4.2 after the bandpass filtering and after the application of a simulated automatic gain control with a cut-off frequency of 300 Hz. In theory, carrier wave amplitude modulations with a frequency of less than $f_l = 300\,\text{Hz}$ should not be caused by the band-limited input signal but by the fading of the channel. An automatic gain control could therefore detect all low-frequency modulations and compensate for all fading with a frequency up to $f_l = 300\,\text{Hz}$. The actual implementations for automatic gain controls and their parameters are however highly manufacturer-specific and very little information is publicly available.

We considered throughout this appendix the case of a small general aviation aircraft as used in the channel measurements presented in Chapter 8. It is also of interest to briefly discuss the case of large commercial aviation passenger aircraft. The relevant parameters are a typical cruising speed of around 250 m/s and a cruising altitude of around 10000 m. In combination with the parameters of Section B.4.1 this results in a maximum Doppler frequency of $f_{D,max} = 100\,\text{Hz}$. The upper band limit of the

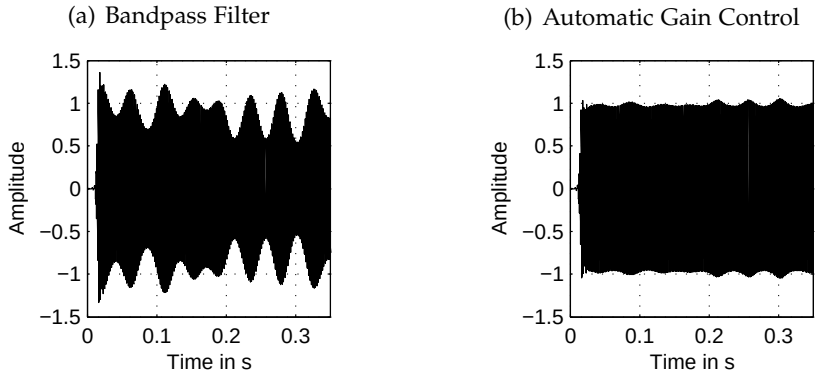


Figure B.13.: Received signal (a) before and (b) after an automatic gain control.

amplitude fading, equaling $2f_{D,max} = 200\,\text{Hz}$, approaches the lower band limit of the audio signal. The path delay of $\Delta l = 6.0\,\text{m}$ or $\Delta\tau = 20\,\text{ns}$ results in a coherence bandwidth of $B_{CB} = 5\,\text{MHz}$, and the channel is thus again frequency nonselective and flat fading.

In this appendix, we dealt with the Doppler shift of the modulated carrier wave, which leads in combination with the multipath propagation to signal fading. It is worthwhile to note that also the modulating signal, that means the audio signal that is represented in the amplitude of the carrier wave, undergoes a Doppler shift. However, this Doppler shift is so small that it can be neglected in practice. For example, for a sinusoidal audio signal with a frequency of 1000 Hz and an aircraft speed of 250 m/s the maximum Doppler frequency using (B.4) is $8.3 \cdot 10^{-4}\,\text{Hz}$.

B.6. Conclusions

We reviewed the most common radio propagation channel models and performed simulations for the aeronautical voice radio channel. We conclude that due to its narrow bandwidth the aeronautical voice radio channel is a frequency nonselective flat fading channel. In most situations the frequency of the amplitude fading is band-limited to the maximum Doppler frequency, which is a simple function of the aircraft speed and the carrier frequency. The amplitude fading of the channel might in parts be mitigated by an automatic gain control in the radio receiver.

Appendix C

Complementary Figures

This appendix presents complementary figures for Section 10.4, which are provided for reference.

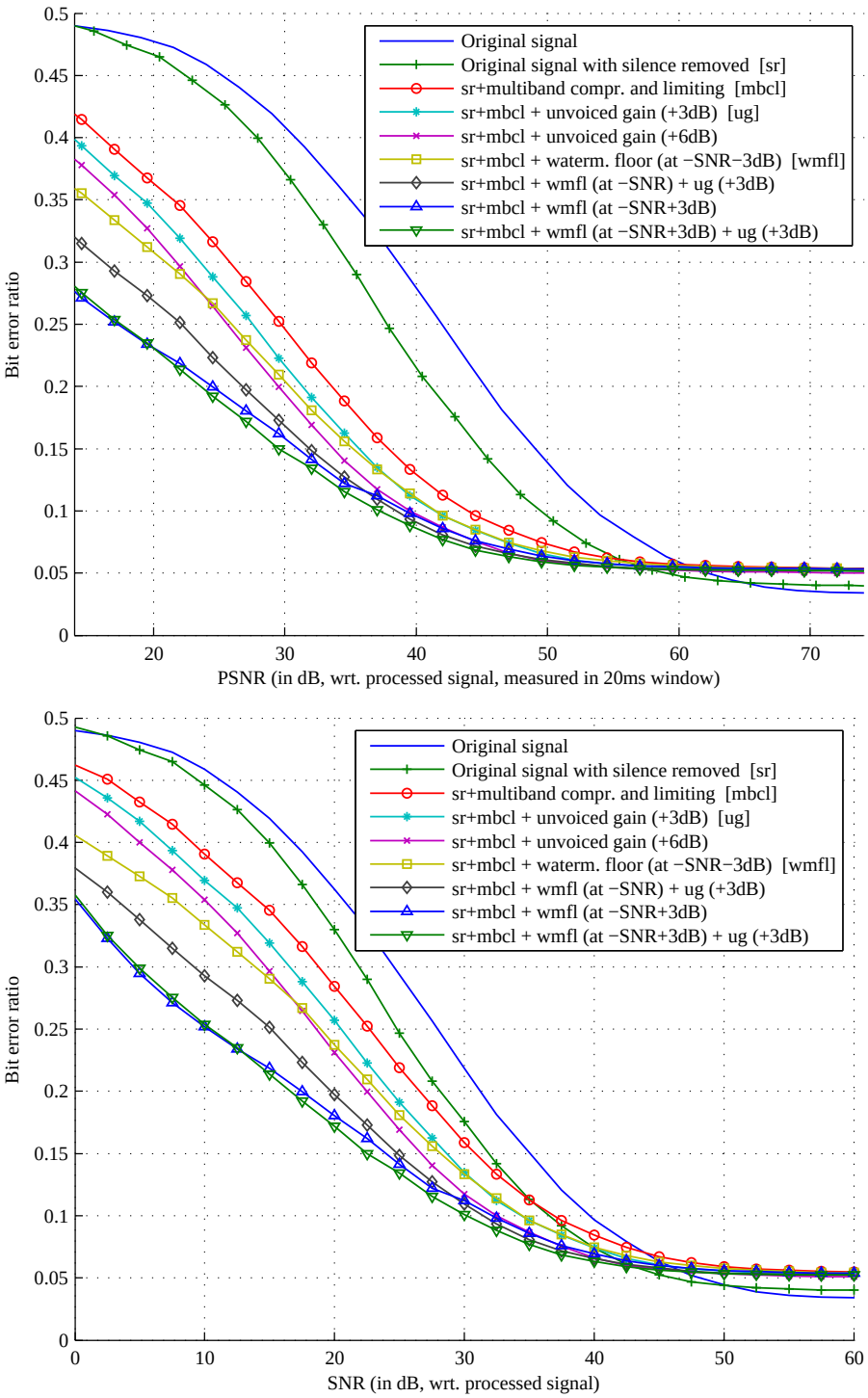


Figure C.1.: Noise robustness curves identical to Figure 10.4, but plotted against different SNR measures.

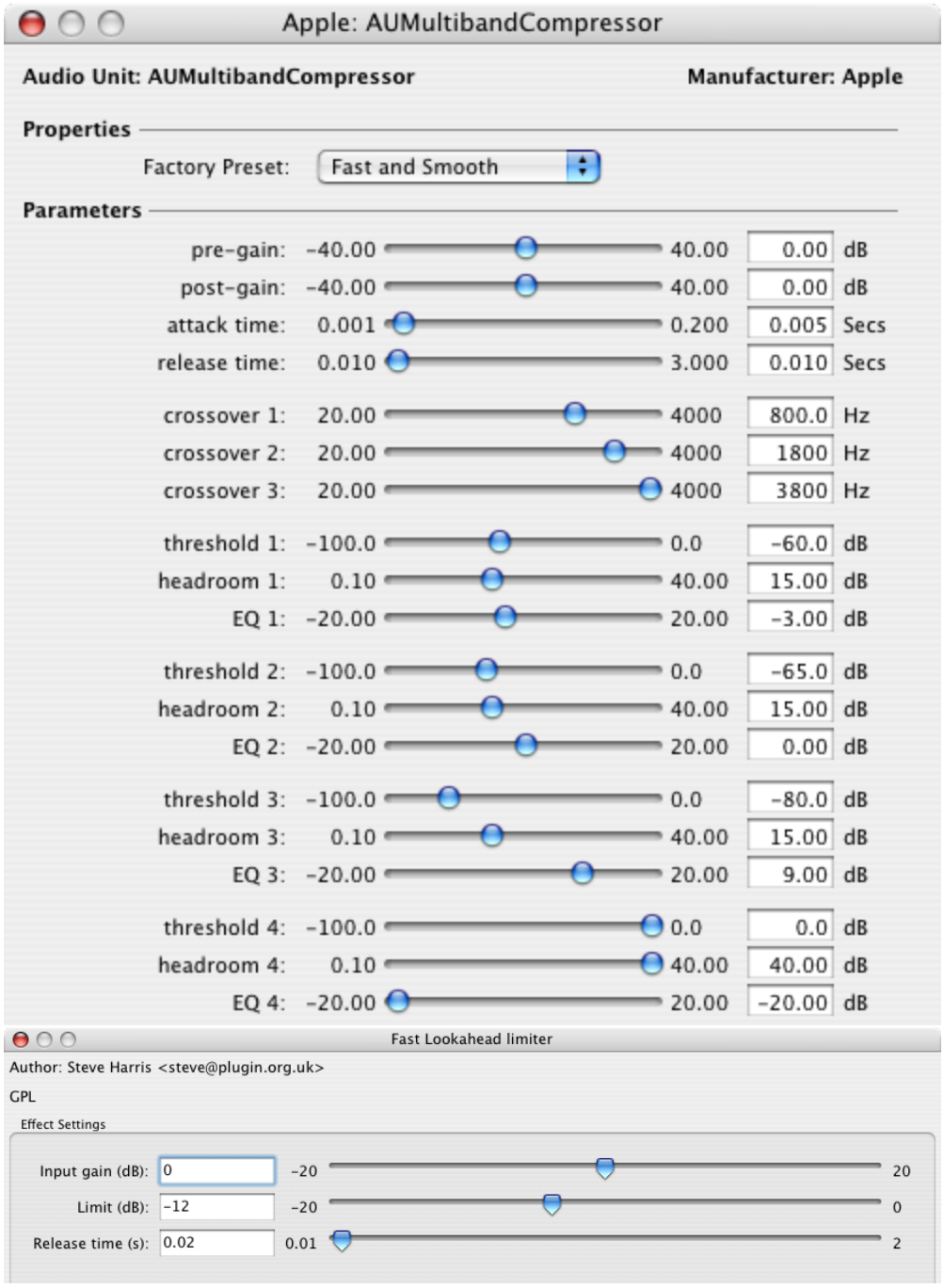


Figure C.2.: Multiband compressor and limiter settings for transmitter-side speech enhancement.

Bibliography

[1] K. Hofbauer, G. Kubin, and W. B. Kleijn, "Speech watermarking for analog flat-fading bandpass channels," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009, revised and resubmitted.

[2] K. Hofbauer and H. Hering, "Noise robust speech watermarking with bit synchronisation for the aeronautical radio," in *Information Hiding*, ser. Lecture Notes in Computer Science. Springer, 2007, vol. 4567/2008, pp. 252–266.

[3] K. Hofbauer and G. Kubin, "High-rate data embedding in unvoiced speech," in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, Pittsburgh, PY, USA, Sep. 2006, pp. 241–244.

[4] K. Hofbauer, H. Hering, and G. Kubin, "A measurement system and the TUG-EEC-Channels database for the aeronautical voice radio," in *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Montreal, Canada, Sep. 2006, pp. 1–5.

[5] K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM corpus of non-prompted clean air traffic control speech," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008.

[6] K. Hofbauer and G. Kubin, "Aeronautical voice radio channel modelling and simulation—a tutorial review," in *Proceedings of the International Conference on Research in Air Transportation (ICRAT)*, Belgrade, Serbia, Jul. 2006.

[7] K. Hofbauer, H. Hering, and G. Kubin, "Speech watermarking for the VHF radio channel," in *Proceedings of the EUROCONTROL Innovative Research Workshop (INO)*, Brétigny-sur-Orge, France, Dec. 2005, pp. 215–220.

[8] K. Hofbauer and H. Hering, "Digital signatures for the analogue radio," in *Proceedings of the NASA Integrated Communications Navigation and Surveillance Conference (ICNS)*, Fairfax, VA, USA, 2005.

[9] K. Hofbauer, "Advanced speech watermarking for secure aircraft identification," in *Proceedings of the EUROCONTROL Innovative Research Workshop (INO)*, Brétigny-sur-Orge, France, Dec. 2004.

[10] M. Gruber and K. Hofbauer, "A comparison of estimation methods for the VHF voice radio channel," in *Proceedings of the CEAS European Air and Space Conference (Deutscher Luft- und Raumfahrtkongress)*, Berlin, Germany, Sep. 2007.

[11] H. Hering and K. Hofbauer, "From analogue broadcast radio towards end-to-end communication," in *Proceedings of the AIAA Aviation Technology, Integration, and Operations Conference (ATIO)*, Anchorage, Alaska, USA, Sep. 2008.

[12] H. Hering and K. Hofbauer, "Towards selective addressing of aircraft with voice radio watermarks," in *Proceedings of the AIAA Aviation Technology, Integration, and Operations Conference (ATIO)*, Belfast, Northern Ireland, Sep. 2007.

[13] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*, 2nd ed. Morgan Kaufmann, 2007.

[14] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3/4, pp. 313–336, 1996.

[15] N. Cvejic and T. Seppanen, *Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks*. IGI Global, 2007.

[16] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[17] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1020–1033, Apr. 2003.

[18] N. Lazic and P. Aarabi, "Communication over an acoustic channel using data hiding techniques," *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 918–924, Oct. 2006.

[19] X. He and M. S. Scordilis, "Efficiently synchronized spread-spectrum audio watermarking with improved psychoacoustic model," *Research Letters in Signal Processing*, vol. 8, no. 1, pp. 1–5, Jan. 2008.

[20] M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[21] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1127–1141, Jul. 1999.

[22] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[23] J. J. Eggers, R. Bäuml, R. Tzschoppe, and B. Girod, "Scalar Costa scheme for information embedding," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1003–1019, Apr. 2003.

[24] F. Perez-Gonzalez, C. Mosquera, M. Barni, and A. Abrardo, "Rational dither modulation: A high-rate data-hiding method invariant to gain attacks," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3960–3975, Oct. 2005.

[25] I. D. Shterev, "Quantization-based watermarking: Methods for amplitude scale estimation, security, and linear filtering invariance," Ph.D. dissertation, Delft University of Technology, 2007.

[26] X. Wang, W. Qi, and P. Niu, "A new adaptive digital audio watermarking based on support vector regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2270–2277, Nov. 2007.

[27] Q. Cheng and J. Sorenson, "Spread spectrum signaling for speech watermarking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Salt Lake City, UT, USA, May 2001, pp. 1337–1340.

[28] M. Hagmüller, H. Hering, A. Kröpfl, and G. Kubin, "Speech watermarking for air traffic control," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, Sep. 2004, pp. 1653–1656.

[29] M. Hatada, T. Sakai, N. Komatsu, and Y. Yamazaki, "Digital watermarking based on process of speech production," in *Multimedia Systems and Applications V*, ser. Proceedings of SPIE, 2002, vol. 4861, pp. 258–267.

[30] S. Chen and H. Leung, "Speech bandwidth extension by data hiding and phonetic classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, Hawaii, USA, Apr. 2007, pp. 593–596.

[31] M. Celik, G. Sharma, and A. M. Tekalp, "Pitch and duration modification for speech watermarking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Philadelphia, PA, USA, Mar. 2005, pp. 17–20.

[32] A. Sagi and D. Malah, "Bandwidth extension of telephone speech aided by data embedding," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 64921, 2007.

[33] B. Geiser, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Lisbon, Portugal, Sep. 2005, pp. 1497–1500.

[34] S. Sakaguchi, T. Arai, and Y. Murahara, "The effect of polarity inversion of speech on human perception and data hiding as an application," in *Proceedings of the*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, Istanbul, Turkey, Jun. 2000, pp. 917–920.

[35] L. Girin and S. Marchand, "Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Montreal, Canada, May 2004, pp. 633–636.

[36] Y.-W. Liu and J. O. Smith, "Audio watermarking through deterministic plus stochastic signal decomposition," *EURASIP Journal on Information Security*, vol. 2007, no. 1, pp. 1–12, 2007.

[37] A. Gurijala and J. Deller, "On the robustness of parametric watermarking of speech," in *Multimedia Content Analysis and Mining*, ser. Lecture Notes in Computer Science. Springer, 2007, vol. 4577/2007, pp. 501–510.

[38] R. C. F. Tucker and P. S. J. Brittan, "Method for watermarking data," U.S. Patent US 2003/0 028 381 A1, Feb. 6, 2003.

[39] S. Chen and H. Leung, "Concurrent data transmission through PSTN by CDMA," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, Island of Kos, Greece, May 2006, pp. 3001–3004.

[40] S. Chen, H. Leung, and H. Ding, "Telephony speech enhancement by data hiding," *IEEE Transactions on Instrumentation and Measurement*, vol. 56, no. 1, pp. 63–74, Feb. 2007.

[41] R. Gray, A. Buzo, A. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 367–376, Aug. 1980.

[42] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," in *Proceedings of the IEEE Workshop on Speech Coding for Telecommunications*, Saint-Adele, Canada, Oct. 1993, pp. 35–36.

[43] D. Schulz, "Improving audio codecs by noise substitution," *Journal of the Audio Engineering Society*, vol. 44, no. 7/8, pp. 593–598, Jul. 1996.

[44] A. Takahashi, R. Nishimura, and Y. Suzuki, "Multiple watermarks for stereo audio signals using phase-modulation techniques," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 806–815, Feb. 2005.

[45] H. M. A. Malik, R. Ansari, and A. A. Khokhar, "Robust data hiding in audio using allpass filters," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1296–1304, May 2007.

[46] H. Matsuoka, Y. Nakashima, T. Yoshimura, and T. Kawahara, "Acoustic OFDM: Embedding high bit-rate data in audio," in *Advances in Multimedia Modeling*, ser. Lecture Notes in Computer Science. Springer, 2008, vol. 0302-9743, pp. 498–507.

[47] X. Dong, M. Bocko, and Z. Ignjatovic, "Data hiding via phase manipulation of audio signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, Montreal, Canada, May 2004, pp. 377–380.

[48] P. Y. Liew and M. A. Armand, "Inaudible watermarking via phase manipulation of random frequencies," *Multimedia Tools and Applications*, vol. 35, no. 3, pp. 357–377, Dec. 2007.

[49] H. Pobloth and W. Kleijn, "On phase perception in speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Phoenix, AZ, USA, Mar. 1999, pp. 29–32.

[50] H. Pobloth, "Perceptual and squared error aspects in speech and audio coding," Ph.D. dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden, Nov. 2004.

[51] S. Haykin, *Communication Systems*, 4th ed. Wiley, 2001.

[52] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[53] H. Fastl and E. Zwicker, *Psychoacoustics*, 3rd ed. Springer, 2006.

[54] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 9, pp. 1292–1304, 2005.

[55] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, no. 3, pp. 337–355, 1998.

[56] A. Sagi and D. Malah, "Data embedding in speech signals using perceptual masking," in *Proceedings of the 12th European Signal Processing Conference (EUSIPCO'04)*, Vienna, Austria, Sep. 2004.

[57] S. Chen and H. Leung, "Concurrent data transmission through analog speech channel using data hiding," *IEEE Signal Processing Letters*, vol. 12, no. 8, pp. 581–584, 2005.

[58] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*. Elsevier, 1995.

[59] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

[60] X. Huang, A. Acero, and H.-W. Hon, Eds., *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Pearson, 2001.

[61] R. C. Beattie, A. Zentil, and D. A. Svihovec, "Effects of white noise on the most comfortable level for speech with normal listeners," *Journal of Auditory Research*, vol. 22, no. 1, pp. 71–76, 1982.

[62] H. J. Larson and B. O. Shubert, *Probabilistic models in engineering sciences*. Wiley, 1979, vol. 1.

[63] N. M. Blachman., "Projection of a spherical distribution and its inversion," *IEEE Transactions on Signal Processing*, vol. 39, no. 11, pp. 2544–2547, Nov. 1991.

[64] R. E. Blahut, *Principles and Practice of Information Theory*. Addison-Wesley, 1987.

[65] J. G. Proakis and M. Salehi, *Communication Systems Engineering*, 2nd ed. Prentice-Hall, 2002.

[66] J. Aldis and A. Burr, "The channel capacity of discrete time phase modulation in AWGN," *IEEE Transactions on Information Theory*, vol. 39, no. 1, pp. 184–185, Jan. 1993.

[67] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Transactions on Information Theory*, vol. 28, no. 1, pp. 55–67, Jan. 1982.

[68] R. Padovani, "Signal space channel coding: Codes for multilevel/phase/frequency signals," Ph.D. dissertation, University of Massachusetts (Amherst), 1985.

[69] D.-S. Kim, "Perceptual phase quantization of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 355–364, Jul. 2003.

[70] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 2117–2119.

[71] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175–205, February 1995.

[72] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 6, pp. 969–978, Jun. 1990.

[73] H. S. Malvar, "Extended lapped transforms: properties, applications, and fast algorithms," *IEEE Transactions on Signal Processing*, vol. 40, no. 11, pp. 2703–2714, Nov. 1992.

[74] H. S. Malvar, *Signal Processing with Lapped Transforms*. Artech House, 1992.

[75] S. Shlien, "The modulated lapped transform, its time-varying forms, and its applications to audio coding standards," *IEEE Transactions on Speec and Audio Processing*, vol. 5, no. 4, July 1997.

[76] Y.-P. Lin and P. Vaidyanathan, "A Kaiser window approach for the design of prototype filters of cosine modulated filterbanks," *IEEE Signal Processing Letters*, vol. 5, no. 6, pp. 132–134, Jun. 1998.

[77] P. Boersma and D. Weenink. (2008) PRAAT: doing phonetics by computer (version 5.0.20). [Online]. Available: http://www.praat.org/

[78] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. (1993) TIMIT acoustic-phonetic continuous speech corpus. CD-ROM. Linguistic Data Consortium. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[79] P. Vary and R. Martin, *Digital Speech Transmission*. Wiley, 2006.

[80] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of the Institute of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

[81] A. J. Jerri, "The Shannon sampling theorem—its various extensions and applications: A tutorial review," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1565–1596, Nov. 1977.

[82] R. G. Vaughan, N. L. Scott, N. L. Scott, D. R. White, and D. R. White, "The theory of bandpass sampling," *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 1973–1984, Sep. 1991.

[83] Y. Wu, "A proof on the minimum and permissible sampling rates for the first-order sampling of bandpass signals," *Digital Signal Processing*, vol. 17, no. 4, pp. 848–854, Mar. 2007.

[84] E. Ayanoglu, N. R. Dagdeviren, G. D. Golden, and J. E. Mazo, "An equalizer design technique for the PCM modem: A new modem for the digital public switched network," *IEEE Transactions on Communications*, vol. 46, no. 6, pp. 763–774, Jun. 1998.

[85] J. Yen, "On nonuniform sampling of bandwidth-limited signals," *IRE Transactions on Circuit Theory*, vol. 3, no. 4, pp. 251–257, Dec. 1956.

[86] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital Communication*, 3rd ed. Springer, 2004.

[87] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice-Hall, 2002.

[88] A. Kohlenberg, "Exact interpolation of band-limited functions," *Journal of Applied Physics*, vol. 24, no. 12, pp. 1432–1436, Dec. 1953.

[89] *ITU-T Recommendation P.48: Specification for an Intermediate Reference System*, International Telecommunication Union, Nov. 1988.

[90] *ITU-T Recommendation G.191: Software Tools for Speech and Audio Coding Standardization*, International Telecommunication Union, Sep. 2005.

[91] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[92] *ITU-T Recommendation P.862.x: Perceptual Evaluation of Speech Quality (PESQ)*, International Telecommunication Union, Oct. 2007.

[93] K. Hofbauer. (2008) Demonstration files: Original and watermarked ATC speech. [Online]. Available: http://www.spsc.tugraz.at/people/hofbauer/wmdemo1/

[94] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 3, pp. 283–296, Jun. 1975.

[95] M. Nilsson, B. Resch, M.-Y. Kim, and W. B. Kleijn, "A canonical representation of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, Hawaii, USA, Apr. 2007, pp. 849–852.

[96] F. Gardner and W. Lindsey, "Guest editorial: Special issue on synchronization," *IEEE Transactions on Communications*, vol. 28, no. 8, pp. 1105–1106, Aug. 1980.

[97] *Airborne VHF Communications Transceiver*, Aeronautical Radio Inc. (ARINC) Characteristic 716-11, Jun. 2003.

[98] L. E. Franks, "Carrier and bit synchronization in data communication–a tutorial review," *IEEE Transactions on Communications*, vol. 28, no. 8, pp. 1107–1121, Aug. 1980.

[99] P. Moulin and R. Koetter, "Data-hiding codes," *Proceedings of the IEEE*, vol. 93, no. 12, pp. 2083–2126, Dec. 2005.

[100] G. Sharma and D. J. Coumou, "Watermark synchronization: Perspectives and a new paradigm," in *Proceedings of the IEEE Conference on Information Sciences and Systems*, Princeton, NJ, USA, Mar. 2006, pp. 1182–1187.

[101] D. J. Coumou and G. Sharma, "Watermark synchronization for feature-based embedding: Application to speech," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Toronto, Canada, Jul. 2006, pp. 849–852.

[102] M. Schlauweg, D. Pröfrock, and E. Müller, "Soft feature-based watermark decoding with insertion/deletion correction," in *Information Hiding*, ser. Lecture Notes in Computer Science. Springer, 2007, vol. 4567/2008, pp. 237–251.

[103] R. A. Scholtz, "Frame synchronization techniques," *IEEE Transactions on Communications*, vol. 28, no. 8, pp. 1204–1213, Aug. 1980.

[104] H. Meyr and G. Ascheid, *Synchronization in Digital Communications*. Wiley, 1990, vol. 1.

[105] F. M. Gardner, *Phaselock Techniques*, 3rd ed. Wiley, 2005.

[106] Y. R. Shayan and T. Le-Ngoc, "All digital phase-locked loop: concepts, design and applications," *IEE Proceedings F: Radar and Signal Processing*, vol. 136, no. 1, pp. 53–56, Feb. 1989.

[107] B. Kim, "Dual-loop DPLL gear-shifting algorithm for fast synchronization," *IEEE Transactions on Circuits and Systems—Part II: Analog and Digital Signal Processing*, vol. 44, no. 7, pp. 577–586, Jul. 1997.

[108] P. A. V. Hall and G. R. Dowling, "Approximate string matching," *ACM Computing Surveys*, vol. 12, no. 4, pp. 381–402, Dec. 1980.

[109] G. Ungerboeck, "Fractional tap-spacing equalizer and consequences for clock recovery in data modems," *IEEE Transactions on Communications*, vol. 24, no. 8, pp. 856–864, Aug. 1976.

[110] D. Artman, S. Chari, and R. Gooch, "Joint equalization and timing recovery in a fractionally-spaced equalizer," in *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, vol. 1, Pacific Grove, CA, USA, Oct. 1992, pp. 25–29.

[111] D. van Roosbroek, *EATMP Communications Strategy - Volume 2 - Technical Description*, 6th ed. EUROCONTROL EATMP Infocentre, 2006.

[112] R. Kerczewski, J. Budinger, and T. Gilbert, "Technology assessment results of the Eurocontrol/FAA future communications study," in *Proceedings of the IEEE Aerospace Conference*, Big Sky, MT, USA, Mar. 2008, pp. 1–13.

[113] *Manual of Radiotelephony*, International Civil Aviation Organization Doc 9432 (AN/925), Rev. 3, 2006.

[114] H. Hering, M. Hagmüller, and G. Kubin, "Safety and security increase for air traffic management through unnoticeable watermark aircraft identification tag transmitted with the VHF voice communication," in *Proceedings of the 22nd Digital Avionics Systems Conference (DASC 2003)*, Indianapolis, USA, 2003.

[115] H. Hering and K. Hofbauer, "System architecture of the onboard aircraft identification tag (AIT) system," Eurocontrol Experimental Centre, EEC Note 04/05, 2005.

[116] M. Celiktin and E. Petre, *AIT Initial Feasibility Study (D1–D5)*. EUROCONTROL EATMP Infocentre, 2006.

[117] J. B. Metzger, A. Stutz, and B. Kauffman, *ARINC Voice Services Operating Procedures Handbook*. Aeronautical Radio Inc. (ARINC), Apr. 2007, rev. S.

[118] M. Sajatovic, J. Prinz, and A. Kroepfl, "Increasing the safety of the ATC voice communications by using in-band messaging," in *Proceedings of the Digital Avionics Systems Conference (DASC)*, vol. 1, Indianapolis, IN, USA, Oct. 2003, pp. 4.E.1–8.

[119] M. Hagmüller and G. Kubin, "Speech watermarking for air traffic control," Eurocontrol Experimental Centre, EEC Note 05/05, 2005.

[120] K. Hofbauer and S. Petrik, "ATCOSIM air traffic control simulation speech corpus," Graz University of Technology, Tech. Rep. TUG-SPSC-2007-11, May 2008. [Online]. Available: http://www.spsc.tugraz.at/ATCOSIM

[121] J. J. Godfrey. (1994) Air traffic control complete. CD-ROM. Linguistic Data Consortium. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S14A

[122] J. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P.-A. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos. (2007) The HIWIRE database, a noisy and non-native english speech corpus for cockpit communication. DVD-ROM. [Online]. Available: http://www.hiwire.org/

[123] S. Pigeon, W. Shen, and D. van Leeuwen, "Design and characterization of the non-native military air traffic communications database (nnMATC)," in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, Antwerp, Belgium, 2007.

[124] L. Graglia, B. Favennec, and A. Arnoux, "Vocalise: Assessing the impact of data link technology on the R/T channel," in *Proceedings of the Digital Avionics Systems Conference (DASC)*, Washington D.C., USA, 2005.

[125] C. Arnoux, L. Graglia, and D. Pavet. (2005) VOCALISE - the today use of VHF as a media for pilots/controllers communications. [Online]. Available: http://www.cena.aviation-civile.gouv.fr/divisions/ICS/projets/vocalise/index_en.html

[126] L. Benarousse, E. Geoffrois, J. Grieco, R. Series, H. Steeneken, H. Stumpf, C. Swail, and D. Thiel, "The NATO native and non-native (N4) speech corpus," in *Proceedings of the RTO Workshop on Multilingual Speech and Language Processing*, no. RTO-MP-066, Aalborg, Denmark, Sep. 2001, pp. 1.1–1.3.

[127] K. Maeda, S. Bird, X. Ma, and H. Lee, "Creating annotation tools with the annotation graph toolkit," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Paris, France, 2002.

[128] C. Mallett. (2008) Autohotkey - free mouse and keyboard macro program with hotkeys and autotext. Computer program. [Online]. Available: http://www.autohotkey.com/

[129] *Designators for Aircraft Operating Agencies, Aeronautical Authorities and Services*, International Civil Aviation Organization Nr. 8585/93, 1994.

[130] F. Schiel and C. Draxler, *Production and Validation of Speech Corpora*. Bastard Verlag, 2003.

[131] A. Campos Domínguez, "Pre-processing of speech signals for noisy and band-limited channels," Master's thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, Mar. 2009, unpublished.

[132] M. Pätzold, *Mobile Fading Channels. Modelling, Analysis and Simulation.* Wiley, 2002.

[133] J. D. Parsons, *The Mobile Radio Propagation Channel*. Wiley, 2000.

[134] BAE Systems Operations, "Literature review on terrestrial broadband VHF radio channel models," B-VHF, Deliverable D-15, 2005. [Online]. Available: http://www.B-VHF.org

[135] D. W. Allan, N. Ashby, and C. C. Hodge, "The science of timekeeping," Agilent Technologies, Application Note AN 1289, 2000.

[136] *MicroTrack 24/96 User Guide*, M-Audio, 2005. [Online]. Available: http://www.m-audio.com

[137] *GPS 35-LVS Technical Specification*, Garmin, 2000. [Online]. Available: http://www.garmin.com

[138] *eTrex Legend Owner's Manual and Reference Guide*, Garmin, 2005. [Online]. Available: http://www.garmin.com

[139] B. Sklar, *Digital Communications*, 2nd ed. Prentice-Hall, 2001.

[140] D. Foster. (2006) GPX: the GPS exchange format. [Online]. Available: http://www.topografix.com/gpx.asp

[141] W. S. Cleveland and S. J. Devlin, "Locally-weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.

[142] M. D. Felder, J. C. Mason, and B. L. Evans, "Efficient dual-tone multifrequency detection using the nonuniform discrete Fourier transform," *IEEE Signal Processing Letters*, vol. 5, no. 7, 1998.

[143] J. O. Smith, *Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications*, 2nd ed. W3K Publishing, 2007.

[144] *MATLAB Version 7.4.0.287 (R2007a)*, The MathWorks, 2007.

[145] M. Gruber, "Channel estimation for the voice radio – Basics for a measurements-based aeronautical voice radio channel model," Master's thesis, FH JOANNEUM - University of Applied Sciences, Graz, Austria, 2007.

[146] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, 1999.

[147] M. J. Levin, "Optimum estimation of impulse response in the presence of noise," *IRE Transactions on Circuit Theory*, vol. 7, pp. 50–56, 1960.

[148] D. D. Rife and J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *Journal of the Audio Engineering Society*, vol. 37, no. 6, pp. 419–444, 1989.

[149] J. Schwaner. (1995) Alternator induced radio noise. Sacramento Sky Ranch Inc. [Online]. Available: http://www.sacskyranch.com/altnoise.htm

[150] *Series 200 Single-Channel Communication System*, Rohde & Schwarz, Munich, Germany, data sheet.

[151] C. Bagwell. (2008) SoX - Sound eXchange (version 13.0.0). [Online]. Available: http://sox.sourceforge.net/

[152] U. Zölzer, *DAFX: Digital Audio Effects*. Wiley, 2002.

[153] D. Mazzoni and R. Dannenberg. (2008) Audacity: The free, cross-platform sound editor (version1.3.6). [Online]. Available: http://audacity.sourceforge.net/

[154] *Core Audio Overview*, Apple, Jan. 2007. [Online]. Available: http://developer.apple.com/documentation/MusicAudio/Conceptual/CoreAudioOverview/

[155] S. Harris. (2006) Steve Harris' LADSPA plugins (swh-plugins-0.4.15). [Online]. Available: http://plugin.org.uk/

[156] *Designators for Aircraft Operating Agencies, Aeronautical Authorities and Services*, International Civil Aviation Organization Nr. 8585/138, 2006.

[157] *Designators for Aircraft Operating Agencies, Aeronautical Authorities and Services*, International Civil Aviation Organization Nr. 8585/107, 1998.

[158] R. Lane, R. Deransy, and D. Seeger, "3rd continental RVSM real-time simulation," Eurocontrol Experimental Centre, EEC Report 315, 1997.

[159] *Digital Aeronautical Flight Information File (DAFIF)*, 6th ed. National Geospatial-Intelligence Agency, Oct. 2006, no. 0610, electronic database.

[160] *Location Indicators*, International Civil Aviation Organization Nr. 7910/122, 2006.

[161] D. Seeger and H. O'Connor, "S08 ANT-RVSM 3rd continental real-time simulation pilot handbook," Dec. 1996, Eurocontrol internal document.

[162] E. Haas. Communications systems. [Online]. Available: http://www.kn-s.dlr.de/People/Haas/

[163] E. Haas, "Aeronautical channel modeling," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 2, pp. 254–264, 2002.

[164] *MATLAB Communications Toolbox*, 3rd ed., The MathWorks, 2004.

[165] K. Metzger. The generic channel simulator. [Online]. Available: http://www.eecs.umich.edu/genchansim/

[166] K. Hofbauer. The generic channel simulator. [Online]. Available: http://www.spsc.tugraz.at/people/hofbauer/gcs/

[167] M. Dickreiter, *Handbuch der Tonstudiotechnik*. KG Saur, 1997, vol. 1.