LECTURE NOTES IN ELECTRICAL ENGINEERING | 13

Bernd Iser
Wolfgang Minker
Gerhard Schmidt

# Bandwidth Extension of Speech Signals



Springer

Bernd Iser • Wolfgang Minker • Gerhard Schmidt

# Bandwidth Extension of Speech Signals

Springer

Bernd Iser
Harman/Becker
Automotive Systems
Söflinger Str. 100
89077 Ulm
Germany
BIser@harmanbecker.com

Gerhard Schmidt
Harman/Becker
Automotive Systems
Söflinger Str. 100
89077 Ulm
Germany
GeSchmidt@harmanbecker.com

Wolfgang Minker
Institute of Information Technology
University of Ulm
Albert-Einstein-Allee 43
89081 Ulm
Germany
wolfgang.minker@uni-ulm.de

ISSN: 1876-1100
ISBN: 978-0-387-68898-5
DOI: 10.1007/978-0-387-68899-2

e-ISSN: 1876-1119
e-ISBN: 978-0-387-68899-2

Library of Congress Control Number: 2008927730



Printed on acid-free paper

springer.com

# Bandwidth Extension of Speech Signals

## Lecture Notes in Electrical Engineering

Volume 13

Bandwidth Extension of Speech Signals
Iser, B.; Minker, Wolfgang; Schmidt, Gerhard
2008, Approx. 200 p., Hardcover
ISBN: 978-0-387-68898-5, Vol. 13

Proceedings of Light-Activated Tissue Regeneration
and Therapy II Conference
Waynant, Ronald; Tata, Darrell B.
2008, Approx. 400 p., Hardcover
ISBN: 978-0-387-71808-8, Vol. 12

Advances in Numerical Methods
Mastorakis, Nikos; Sakellaris, John (Eds.)
2008, Approx. 300 p., Hardcover
ISBN: 978-0-387-76482-5, Vol. 11

Embedded Systems Specification and Design Languages
Villar, Eugenio (Ed.)
2008, Approx. 400 p., Hardcover
ISBN: 978-1-4020-8296-2, Vol. 10

Content Delivery Networks
Buyya, Rajkumar; Pathan, Mukaddim; Vakali, Athena (Eds.)
2008, Approx. 400 p., Hardcover
ISBN: 978-3-540-77886-8, Vol. 9

Unifying Perspectives in Computational and Robot Vision
Kragic, Danica; Kyrki, Ville (Eds.)
2008, Approx. 250 p., Hardcover
ISBN: 978-0-387-75521-2, Vol. 8

Sensor and Ad-Hoc Networks
Makki, S.K.; Li, X.-Y.; Pissinou, N.; Makki, S.; Karimi, M.; Makki, K. (Eds.)
2008, Approx. 350 p. 20 illus., Hardcover
ISBN: 978-0-387-77319-3, Vol. 7

Trends in Intelligent Systems and Computer Engineering
Castillo, Oscar; Xu, Li; Ao, Sio-Iong (Eds.)
2008, Approx. 750 p., Hardcover
ISBN: 978-0-387-74934-1, Vol. 6

(*continues after index*)

# Preface

The idea for this book was formed during the doctorate of Bernd Iser. Bernd Iser was working on efficient and robust bandwidth extension algorithms in hands-free systems for Harman/Becker Automotive Systems. It turned out that bandwidth extension of speech signals was a topic of appreciable interest, where lots of scientific publications discussing details of specific solutions could be found. What was missing was a contribution elaborately discussing the entirety of different approaches and comparing, respectively evaluating them in a meaningful manner. Another property that was disregarded in the state of the art was the influence of noise corrupted real-world signals. All these considerations led to the belief that there was a need for a book taking all these missing aspects into account.

Prof. Dr. Wolfgang Minker, who was supervising the doctorate of Bernd Iser at the University of Ulm, Germany, came up with the idea of writing such a book. Dr. Gerhard Schmidt, the leader of the acoustic signal processing research team, Bernd Iser was working at Harman/Becker Automotive Systems, joined the project for having another speech signal processing expert on board. Based on the research work on the topic of bandwidth extension of speech signals the present book emerged in collaboration that tries to cover the above described requirements.

Like all extensive projects this book project would not have been possible to handle without the support, advice, criticism and review of countless people. Representative for all of them a few people that added substantial contributions to this book should be mentioned and thanked for their valuable support.

The authors thank Dr. Markus Buck for his accurate and detailed review of the manuscript; Mohamed Krini for many valuable discussions on the algorithms as well as on the manuscript and Dr. Tim Haulick for arranging the possibility of conducting research on the topic of bandwidth extension of speech signals at Harman/Becker Automotive Systems.

*Bernd Iser*
*Wolfgang Minker*
*Gerhard Schmidt*

(*continued from page ii*)

Advances in Industrial Engineering and Operations Research
Chan, Alan H.S.; Ao, Sio-Iong (Eds.)
2008, XXVIII, 500 p., Hardcover
ISBN: 978-0-387-74903-7, Vol. 5

Advances in Communication Systems and Electrical Engineering
Huang, Xu; Chen, Yuh-Shyan; Ao, Sio-Iong (Eds.)
2008, Approx. 700 p., Hardcover
ISBN: 978-0-387-74937-2, Vol. 4

Time-Domain Beamforming and Blind Source Separation
Bourgeois, J.; Minker, W.
2009, Approx. 200 p., Hardcover
ISBN: 978-0-387-68835-0, Vol. 3

Digital Noise Monitoring of Defect Origin
Aliev, T.
2007, XIV, 223 p. 15 illus., Hardcover
ISBN: 978-0-387-71753-1, Vol. 2

Multi-Carrier Spread Spectrum 2007
Plass, S.; Dammann, A.; Kaiser, S.; Fazel, K. (Eds.)
2007, X, 106 p., Hardcover
ISBN: 978-1-4020-6128-8, Vol. 1

Printed in the United States of America

# Contents

# 1

# Introduction

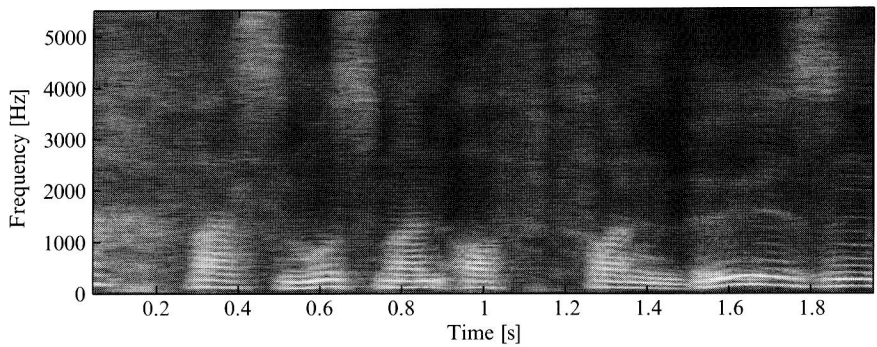
Speech is a natural and therefore privileged communication modality. This is the reason for the great success of speech driven services and speech based media. Multimedia would not be imaginable without high quality audio. Today's environment is full of high quality audio sources like CD-Audio, DVD-Audio, radio broadcast, television broadcast, and so on.

One of the oldest but still most popular media based on audio is the telephone network. But since its invention in the nineteenth century capabilities and simultaneously demands on audio quality have increased [Bell 77]. Today's telephone networks still provide poor audio quality due to historical limitations (see Sect. 1.3).

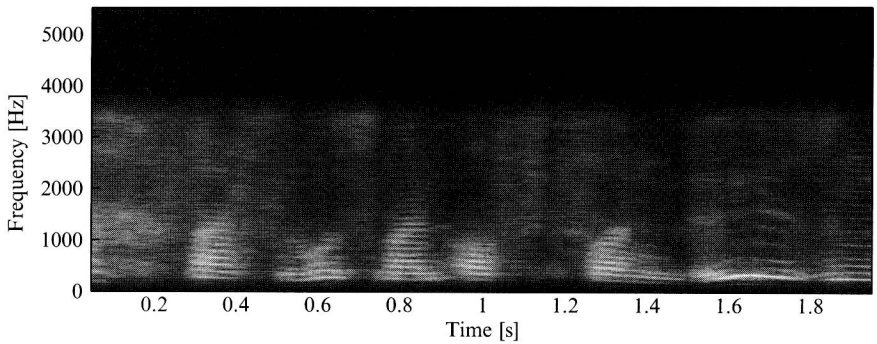
The reason for the poor audio quality of the telephone network is the very limited bandwidth that is provided. Analog networks, for example, provide only a bandwidth of about 3.1 kHz [ITU 88b] (see Fig. 1.1b). This leads to reduced speech quality and even intelligibility. A typical property of analog networks is the difficulty of distinguishing between several fricatives like present in the words "feel" ([fi:l][1]) and "veal" ([vi:l]). Another typical problem is that one is not able to distinguish similar voices (father–son problem) over the telephone.

By using more modern media this drawback becomes even more obvious. This can be best experienced by listening to radio or CD within a car and afterwards using the hands-free system. Another example are the telephone services that are recently available in the internet like skype™, which have a considerably higher bandwidth than conventional telephone networks. However the telephone network is still one of the most widespread networks all over the world. This deed has made attempts to change the network in order to provide a better audio quality doomed to failure due to the massive effort of exchanging the hardware.

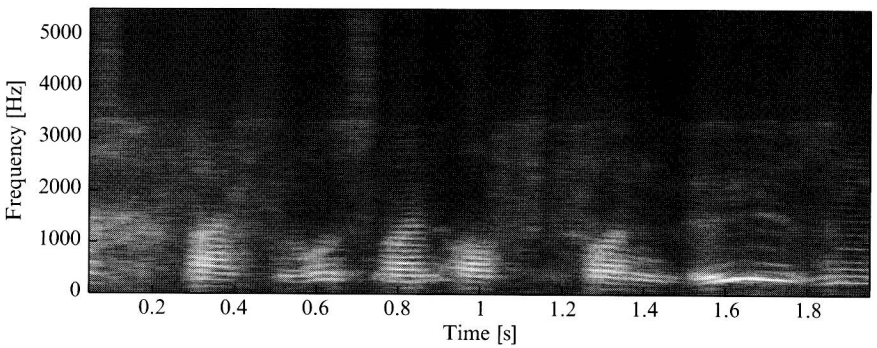
[1] Phonetic description according to [IPA 49].




(a) Wideband speech.




(b) Band limited speech.




(c) Reconstructed speech.

**Fig. 1.1.** Spectrograms of (**a**) wideband speech, (**b**) band limited speech, and (**c**) reconstructed speech. The spectrograms (**a**) and (**c**) are limited by the sampling rate of 11,025 Hz. The spectrogram in (**b**) is limited by an analog telephone bandpass with a passband from approximately 300 to 3,400 Hz causing a reduced speech quality in comparison to (**a**) which could be a recording close to the mouth with no attenuation. The spectrogram (**c**) contains the estimates for the frequency components missing in (**b**). Methods for estimating these components will be subject of this work

This is the point where the idea of bandwidth extension comes into mind quite intuitively. Bandwidth extension in this context means the estimation of the not transmitted frequency components out of the transmitted signal by exploiting the transinformation included in speech signals and therewith increasing the speech quality (see Fig. 1.1c). This approach yields the advantage that nothing has to be changed within the network – it is simply an optional feature on the terminal side.

## 1.1 Scope of the Book

This book is intended to provide a profound knowledge on the problem of bandwidth extension of band limited speech signals in a first step. Theory and methods for quality enhancement of speech signals that have undergone a bandlimitation, as it is the case for example in a telephone network, will be described. These enhancements have been performed with and without the presence of noise. In a second step, novel, not yet published solutions as well as improvements for the problem of bandwidth extension will be presented. In contrast to prior work emphasis is placed on novel approaches to handle problems that emerge from dealing with real-world signals and environments comprising all kinds of disturbances. Another previously not considered aspect is the subjective evaluation and comparison of the different methods. By analyzing the usability of well known objective distance measures the need of a measure that better correlates with subjective evaluation resulted in the development of a novel objective distance measure.

All band limited signals that are matter of interest in this book are speech signals that have undergone a band limitation by the application of a telephone bandpass. But the presented algorithms are not restricted to telephone band limited signals. They are also applicable to speech signals that are band limited in any other way as for example by downsampling with prior lowpass filtering. Another field of application would be the re-synthesis of heavily perturbed speech signals (see [Hosoki 02, Seltzer 05b, Yegnanarayana 97]) that can not sufficiently be enhanced in terms of quality by the standard noise reduction systems [Hänsler 03]. The scope of the book will include topics related to speech coding, pattern and speech recognition, speech enhancement, statistics and digital signal processing in general.

The study and development of several methods for the bandwidth extension of speech signals will be described. Problems and the respective solutions are discussed for the different approaches. Since all described methods will be based on the source-filter model, this model will be presented after a short problem motivation and illustration of the process of human speech generation. The proposed methods will include the extraction of speech-model based parameters like cepstral coefficients, auto-regressive parameters (all-pole-filter coefficients), line spectral pairs (LSP) or mel-frequency cepstral coefficients

(MFCC). These parameters and their properties will briefly be presented. Besides these parametric representations of the spectral envelope some other scalar speech features like the zero-crossing rate, gradient index, pitch frequency, local kurtosis, spectral centroid, and energy based features will be introduced. For evaluating the different approaches as well as for the training and operation, a couple of well known spectral distortion measures and a new measure will be presented. After these rather general topics the main algorithms will be introduced. These can be divided into two separate subtasks, namely the generation of a broadband excitation signal and the generation of a broadband spectral envelope. For the first one, several approaches like non-linear characteristics, spectral shifting, signal generators, etc. will be explained. For the latter one, codebook approaches, neural networks, linear mapping, as well as joint approaches will be discussed. For the generation of the broadband spectral envelope the proposed methods need a prior training phase. This part and the speech corpora which have been used as well as the specific demands and preparation of these corpora will be part of this section. Another aspect of this section will be the challenge of facing different transmission paths with their respective frequency responses that have not been part of the training scenario and schemes to overcome this problem. Afterwards focus will be placed on the performance of these methods. For the evaluation of the performance the already introduced objective spectral distortion measure will be used. Additionally, a subjective evaluation will be presented by analyzing the result of listeners voting the different approaches. A significance analysis will also be drawn for the subjective evaluation as well as a rank correlation between the subjective results and the respective objective distortion measures to evaluate the significance of these objective measures. The most promising method has been implemented in a real-time environment using appropriate hardware. This system will be described as well. The book will close with a summary of the described methods and the results of the evaluation.

The speech signals within this book emerge from the car environment and have a sampling rate of 11,025 Hz and a resolution of 16 bit. This is the sampling rate and the resolution that would be available for example when using one subsampled (factor 4) channel of a MOST-bus (Multimedia Oriented System Transfer) as it is installed in many present cars. However, the presented algorithms are neither restricted to this sampling rate nor to the car environment.

## 1.2 Nature of Speech Signals

Human speech generally occupies approximatively the whole frequency range that is perceptible for the auditory system [Zwicker 99]. In Fig. 1.2 the estimate for the power spectral density of a speech sample recorded with a sampling rate of 44.1 kHz is depicted. It is clearly visible that this signal possesses

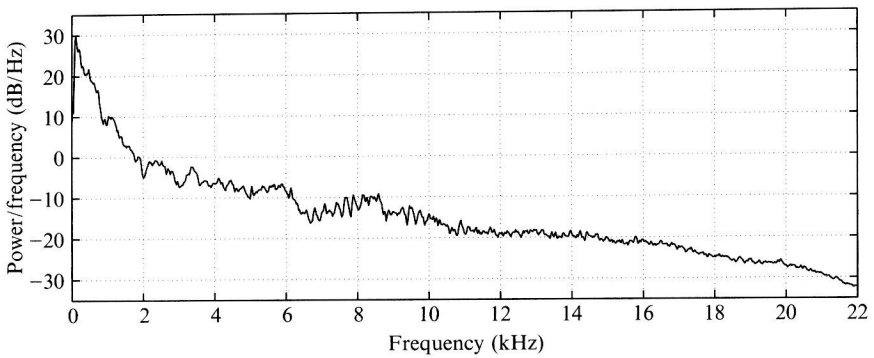
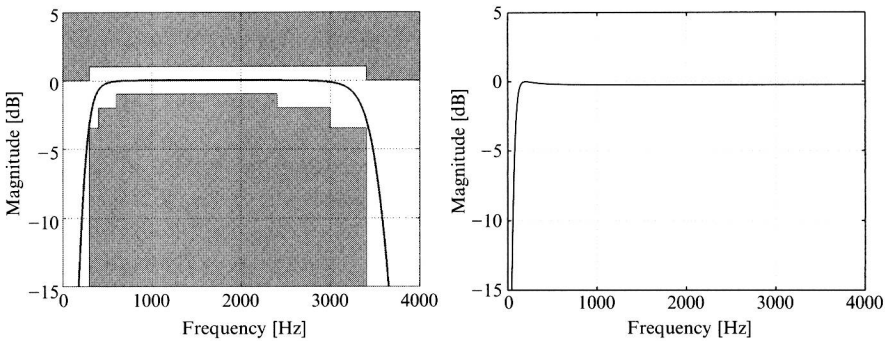



**Fig. 1.2.** Power spectral density estimate of a speech sample (male speaker; approximately 40 s of speech). Note that the signal has not been equalized for the microphone transfer function




(a) Analog. (b) GSM-EFR.

**Fig. 1.3.** (**a**) Telephone bandpass according to [ITU 88b] and (**b**) highpass as implemented in GSM-EFR according to [ETSI 00]

power up to 22.05 kHz. This broad frequency range however is not necessary for speech intelligibility. The speech intelligibility decreases only marginally if band limiting the speech signal by using a sampling rate of 16 kHz for example. However, the speech quality decreases remarkably.

## 1.3 Band Limited Speech Signals

The degradation of speech quality using analog telephone systems is caused by the introduction of band-limiting filters within amplifiers used to keep a certain signal level in long local loops [Kammeyer 92]. These filters have a passband from approximately 300 up to 3,400 Hz (see Fig. 1.3a) and are applied to reduce the crosstalk between different channels.

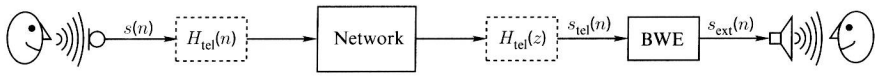



**Fig. 1.4.** Overall system for bandwidth extension. The *dashed line* indicates that the location(s) where the band limitation(s) take(s) place depend(s) on the routing of the call

As can be seen in Fig. 1.1b the application of such a bandpass (depicted in Fig. 1.3a) attenuates considerable speech portions. Digital networks, such as the integrated service digital network (ISDN) and the global system for mobile communication (GSM), are able to transmit speech in higher quality since signal components below 300 Hz as well as components between 3.4 and 4 kHz can be transmitted (see Fig. 1.3b). Considering a transmission over a GSM network the only two bandlimiting factors are the sampling rate of 8 kHz limiting the signal to an upper limit of 4 kHz and the lowpass filter depicted in Fig. 1.3b that is specified for example in the enhanced full-rate coder [ITU 88b]. However, this is only true if the entire call (in terms of its routing) remains in those networks – when leaving into an analog telephone network, the speech signal is once again band limited (see dashed rectangles in Fig. 1.4).

Thus, great efforts have been made to increase the quality of telephone speech signals in recent years. Wideband codecs are able to increase the bandwidth up to 7 kHz or even higher at only moderate complexity [ITU 88a, Croll 72]. Other approaches try to increase the bandwidth by transmitting side information on the missing frequency bands or by combined estimation and coding (see [Agiomyrgiannakis 04, Geiser 05]). Nevertheless, applying these codecs would require an exchange of the current networks or at least, in the second case, the usage of two devices that are able to encode side information using acoustic watermarking on the far end side and to decode such information on the local side. Acoustic watermarking in this context means to hide information in the audio data that is however not audible but still possible to decode. Another possibility is to increase the bandwidth after transmission by means of bandwidth extension (for other approaches, e.g. psychoacoustic approaches, see [Boillot 05]). The basic idea of these enhancements is to estimate the speech signal components above 3400 Hz and below 300 Hz and to complement the signal in the idle frequency bands with this estimate. Precondition is the sufficient correlation between the speech signal in the telephone band and the extension regions which is discussed in [Jax 02b]. In this case the telephone networks remain untouched. Figure 1.4 shows the basic structure of a bandwidth extension (BWE) system included in the receiving path of a telephone connection.

Additionally, three time-frequency analyses are presented in Fig. 1.1. The first analysis depicts a wideband speech signal $s(n)$ recorded close to the mouth of the communication partner on the remote side. If we assume no errors or distortions during the transmission, a bandlimited signal $s_{\mathrm{tel}}(n)$