Statistical Analysis of Cepstral Coefficients and Applications in Speech Enhancement

Dissertation

zur Erlangung des Grades eines
Doktor-Ingenieurs
der Fakultät für Elektrotechnik
und Informationstechnik
an der Ruhr-Universität Bochum

von

Dipl.-Ing. Timo Florian Gerkmann

Bochum 2010

Zusammenfassung

Von digitalen Kommunikationsgeräten, wie Hörhilfen oder Mobiltelefonen, werden häufig Sprachsignale erfasst, die durch Umgebungsgeräusche gestört sind. Mit Hilfe von Geräuschreduktionsalgorithmen kann das Verhältnis zwischen Sprach- und Geräuschsignalleistung (SNR) verbessert werden. Die Verbesserung des SNRs geht allerdings oft mit Prozessierungsartefakten oder Sprachsignalverzerrungen einher. Ziel dieser Arbeit ist es, ohne Einbußen in der Geräuschreduktion Artefakte und Sprachsignalverzerrungen zu reduzieren.

Bei den in dieser Arbeit verwendeten Geräuschreduktionsalgorithmen wird das gestörte Sprachsignal mit einer gleitenden diskreten Fouriertransformation (DFT) in den Spektralbereich transformiert, und dann in Abhängigkeit einer adaptiven Schätzung des a priori SNRs mit einer Gewichtungsfunktion multipliziert. Ist das a priori SNR lokal überschätzt, kommt es zu Ausreißern im prozessierten Sprachsignal, die häufig als tonale Artefakte wahrgenommen werden. Eine Unterschätzung führt hingegen zu Sprachsignalverzerrungen. Das Cepstrum bezeichnet die inverse diskrete Fouriertransformierte des logarithmierten Betragsquadrats einer spektralen Größe, und hat die Eigenschaft, dass spektrale Ausreißer einerseits und spektrale Sprachaktivität andererseits durch überwiegend disjunkten Mengen cepstraler Koeffizienten repräsentiert werden. Daher lassen sich, ohne wesentliche Zunahme der Sprachsignalverzerrungen, durch eine selektive Glättung im Cepstralbereich spektrale Ausreißer effektiv reduzieren.

In dieser Arbeit werden die statistischen Eigenschaften cepstraler Koeffizienten analysiert. Zunächst erfolgt die Herleitung von Gleichungen für den Mittelwert und die Varianz cepstraler Koeffizienten und logarithmierter spektraler Betragsquadrate. Es wird gezeigt, dass eine χ^2-verteilte spektrale Zufallsgröße nach einer cepstralen Glättung noch annähernd χ^2-verteilt ist, allerdings eine erhöhte Zahl Freiheitsgrade aufweist. Es wird eine Gleichung hergeleitet, welche die Bestimmung der Freiheitsgrade nach einer cepstralen Glättung in Abhängigkeit der Glättungsparameter erlaubt. Mit Hilfe dieser Ergebnisse wird nachgewiesen, dass eine erwartungstreue Glättung im Cepstralbereich zu einem systematischen Fehler im Spektralbereich führt. Wenn spektrale Größen wie das a priori SNR im Cepstralbereich geglättet werden, führt dieser systematische Fehler zu einer Unterschätzung der spektralen Gewichtungsfunktion und kann somit Sprachsignalverzerrungen hervorrufen. In dieser Arbeit wird der systematische Fehler in Abhängigkeit der verwendeten Glättungsparameter analytisch beschrieben, und kann somit korrigiert

werden. Dadurch wird eine cepstrale Glättung möglich, die auch im Spektralbereich erwartungstreu ist.

Zur Bestimmung der Untermenge sprachrelevanter Cepstralkoeffizienten ist eine Schätzung der Sprachgrundfrequenz nötig. Es wird hergeleitet, dass hierfür die Suche des Maximums der Cepstralkoeffizienten, welche die spektrale Feinstruktur darstellen, optimal im Maximum-Likelihood Sinn ist. Sind mehrere Mikrofone vorhanden, so erhält man die optimale Lösung, wenn die Cepstren der Mikrofonsignale vor der Maximumssuche addiert werden. Weiterhin kann die Korrelation der Sprachgrundfrequenz zeitlich aufeinanderfolgender Segmente für eine robustere Schätzung im Sinne eines Maximum-A-Posteriori Schätzers genutzt werden.

Es wird belegt, dass eine zeitliche Glättung des Cepstrums spektraler Größen, wie der a priori SNR Schätzung, spektrale Ausreißer reduziert und unter Verwendung des hergeleiteten Korrekturterms im Vergleich zu herkömmlichen Verfahren sowohl zu einer Erhöhung des Ausgangs-SNRs als auch zu weniger Sprachsignalverzerrungen führt. Ähnliche Ergebnisse werden erzielt wenn cepstrale Koeffizienten des prozessierten Signals durch Koeffizienten des unprozessierten Signals ersetzt werden. Der Vorteil der cepstralen Ersetzung im Gegensatz zu einer zeitlichen cepstralen Glättung ist, dass eine zeitliche Verzerrung des prozessierten Signals ausbleibt. Dadurch wird insbesondere bei nichtstationären Geräuschquellen, wie konkurrierenden Sprechern, die Natürlichkeit des prozessierten Signals weiter erhöht. Der Vorteil der zeitlichen cepstralen Glättung ist hingegen ein geringerer Rechenaufwand. Zudem wird gezeigt, dass der Rechenaufwand durch eine Modifikation der Spektraltransformationen, die zur Berechnung des Cepstrums verwendet werden, deutlich reduziert werden kann, während die Qualität des Ausgangssignals annähernd unverändert bleibt.

Abschließend wird die Schätzung der a posteriori Sprachanwesenheitswahrscheinlichkeit (SPP) in jedem spektralen Koeffizienten behandelt. Im Gegensatz zu konkurrierenden Verfahren werden a priori SNR und a priori SPP nicht adaptiert, sondern vorab bestimmt, was zu einer Entkopplung der a posteriori SPP Schätzung und der Schätzung der spektralen Gewichtungsfunktion führt. Der vorgeschlagene Algorithmus führt zu weniger Schätzfehlern als konkurrierende Verfahren. Dies wird besonders dann deutlich, wenn der SPP Schätzer mit einer zeitlichen Glättung des Cepstrums, dem hergeleiteten Korrekturterm und dem vorgeschlagenen Verfahren zur Bestimmung der Freiheitsgrade kombiniert wird.

Contents

Abbreviations

SNR	Signal-to-Noise Ratio
SPP	Speech Presence Probability
TCS	Temporal Cepstrum Smoothing
CN	Cepstral Nulling
MMSE	Minimum Mean Square Error
ML	Maximum Likelihood
MAP	Maximum *A Posteriori*
DFT	Discrete Fourier Transform
IDFT	Inverse Discrete Fourier Transform
FFT	Fast Fourier Transform
DCT	Discrete Cosine Transform
IDCT	Inverse Discrete Cosine Transform
PESQ	Perceptual Evaluation of Speech Quality
MOS	Mean Opinion Score
LSA	Log Spectral Amplitude
GER	Gross Error Rate
RMSE	Root Mean Square Error
MM-CML	Multi-Microphone Cepstral ML fundamental period estimator
BF-CML	Beamforming based Cepstral ML fundamental period estimator
MM-CMAP	Multi-Microphone Cepstral MAP fundamental period estimator
SD	Speech Distortions
NL	Noise Leakage

Symbols

Chapter 1

γ_k	*a posteriori* Signal-to-Noise Ratio (SNR)
$\sigma^2_{\mathrm{N},k}$	Spectral variance of the noise signal
$\sigma^2_{\mathrm{S},k}$	Spectral variance of the speech signal
ξ_k	*a priori* SNR
$\mathrm{E}\{\cdot\}$	Mathematical expectation
G_k	Spectral gain function
$\widetilde{G}_k$	Limited spectral gain function
$G_{\min}$	Lower limit on the spectral gain function
$\mathcal{H}_{0,k}$	Hypothesis that speech is absent
$\mathcal{H}_{1,k}$	Hypothesis that speech is present
k	Frequency index
L	Signal segment advance, undersampling factor
l	Signal segment index
N	Signal segment length
τ	Discrete time index
w_n	Spectral analysis window
$\tilde{w}_n$	Synthesis window
N_k	Noise only DFT coefficients
S_k	Clean speech DFT coefficients

Y_k Noisy DFT coefficients

$y(\tau)$ Noisy discrete time domain signal

Chapter 2

α_q Recursive cepstral smoothing factor

ε_q Cepstral mean deviation

$\Gamma(\cdot)$ Complete gamma function [Gradshteyn and Ryzhik, 2000, (8.31)]

κ_m covariance of a logarithmic periodogram bin and its mth neighbor

μ Shape parameter of a χ or χ^2 distribution, 2μ are the degrees of freedom

$\bar{\mu}$ Shape parameter of a χ or χ^2 distribution after cepstral smoothing

Φ_k Positive, symmetric, real valued spectral quantity such as a periodogram, or a spectral gain function

$\bar{\Phi}_k$ Positive, symmetric, real valued spectral quantity after unbiased cepstral smoothing

$\tilde{\Phi}_k$ Positive, symmetric, real valued spectral quantity after cepstral smoothing without bias correction

ϕ_q Symmetric, real valued cepstral quantity

$\bar{\phi}_q$ Symmetric, real valued cepstral quantity after cepstral smoothing

$\check{\phi}_q$ Cepstral quantity that is convolved with a Hamming window

$\psi(\cdot)$ psi-function [Gradshteyn and Ryzhik, 2000, (8.360)]

ρ_m correlation coefficient of the frequency coefficient S_k and its mth neighbor S_{k+m}

$\zeta(\cdot,\cdot)$ Riemann's zeta-function [Gradshteyn and Ryzhik, 2000, (9.521.1)]

$\mathcal{B}$ Bias correction factor

b_q Indicator function for Cepstral Nulling (CN)

f_s Sampling rate

$\log(\cdot)$ Natural logarithm

P_k Periodogram, $P_k = |S_k|^2$

q Cepstral index

q_0 Fundamental period index

$\mathbb{Q}$ Cepstral Coefficients that represent the speech spectral structure

$\overline{\mathbb{Q}}$ Cepstral Coefficients that are assumed not to represent information on the speech spectral structure

q_{low} The cepstral coefficients $q \leq q_{\text{low}}$ that are assumed to represent the speech spectral envelope

$|S_k|$ Spectral amplitudes

$|\bar{S}_k|$ Unbiased spectral amplitudes after cepstral smoothing

$|\widetilde{S}_k|$ Spectral amplitudes after cepstral smoothing without bias correction

Chapter 3

α_{dd} Recursive smoothing constant of the decision-directed approach

$\delta(\cdot)$ Dirac's delta function

$\Theta(\cdot)$ Heaviside's step function

Chapter 4

μ_{MS} Degrees of freedom after the optimal smoothing proposed in [Martin, 2001]

$Y_{\text{MS},k}$ Noisy observation after the optimal smoothing proposed in [Martin, 2001]

Chapter 5

$\gamma_k^{\text{intersect}}$ Value of γ_k that results in $\Lambda_k = 1$

Λ_k Generalized likelihood ratio

ξ_{fix} Optimal fixed *a priori* SNR

c_f Costs for false-alarms

c_m Costs for missed-hits

P_f False-alarm rate

P_m Missed-hit rate

$\mathcal{R}$ Risk

List of Algorithms

List of Figures

List of Tables

Chapter 1

Introduction

In digital communication devices, such as hearing aids or cellular telephones, speech signals are captured by one or more microphones. The speech signal is often disturbed by additive background noise, such as competing speakers or traffic noise. In the last decades a tremendous progress has been made in the development of speech enhancement algorithms that are capable of reducing additive noise with only little speech distortions. In order to achieve this, single channel speech enhancement algorithms exploit the different statistical properties of speech and noise signals. For slowly varying noise sources, such as interior car noise, state-of-the-art algorithms achieve an evident enhancement in the signal quality. However, in nonstationary noise environments such as babble noise, single channel speech enhancement remains a challenging task. This thesis aims at increasing the robustness of speech enhancement algorithms in nonstationary noise environments.

1.1 Speech enhancement in the short-time discrete Fourier domain

In this thesis, we address speech enhancement algorithms that work in the short-time discrete Fourier domain. For the spectral analysis, segments of the noisy discrete observation $y(\tau)$ are weighted by a window w_n, and transformed into the discrete Fourier domain, as

$$Y_k(l) = \sum_{n=0}^{N-1} w_n \, y(lL+n) \, \mathrm{e}^{-\mathrm{j}2\pi kn/N} \,, \tag{1.1}$$

where τ and n are the discrete time indices, l is the segment index, k is the frequency index, L is the segment shift, and N is the segment size. It is assumed that under speech presence the noisy observation is a linear superposition of clean speech

$S_k(l)$ and noise $N_k(l)$. Thus, the observed signal under the hypothesis $\mathcal{H}_{1,k}(l)$, which signifies the presence of speech, is given as $Y_k(l) = S_k(l) + N_k(l)$. Under hypothesis $\mathcal{H}_{0,k}(l)$ that indicates the absence of speech, the observed signal takes the form $Y_k(l) = N_k(l)$. Subsequently, whenever possible, we omit the frame index l for notational convenience.

The considered speech enhancement algorithms can be decomposed into three blocks:

1. the estimation of the noise spectral power given the noisy observation,
2. the estimation of the speech spectral power given the estimated noise spectral power and the noisy observation,
3. the estimation of clean speech spectral coefficients given the speech spectral power, the noise spectral power, and the noisy observation.

The spectral noise power $\sigma^2_{\mathrm{N},k} = \mathrm{E}\{N_k\}$ can be estimated *e.g.* using the minimum statistics approach [Martin, 2001], minimum controlled recursive averaging [Cohen, 2003], subspace decomposition [Hendriks *et al*, 2008], or Minimum Mean Square Error (MMSE) estimators [Hendriks *et al*, 2010]. For the evaluation of the algorithms presented in this thesis, Martin's minimum statistics approach [Martin, 2001] is used for spectral noise power estimation. This approach is based on the observation that a noise power estimate can be obtained by using minimum values of a power estimate of the noisy signal. As such, it assumes that the noise signal is more stationary than the speech signal, and that the found minimum represents only the noise signal. The noise spectral power is then inferred from the found minima. With the spectral noise power known, the noisy observation is normalized to obtain the *a posteriori* Signal-to-Noise Ratio (SNR)

$$\gamma_k = |Y_k|^2/\sigma^2_{\mathrm{N},k}\,.$$

The speech power $\sigma^2_{\mathrm{S},k} = \mathrm{E}\{|S_k|^2\}$ is often implicitly estimated by using the decision-directed approach [Ephraim and Malah, 1984] that estimates the *a priori* SNR

$$\xi_k = \sigma^2_{\mathrm{S},k}/\sigma^2_{\mathrm{N},k}\,.$$

In Section 3.2 we discuss the *a priori* SNR estimation in more detail and propose a novel, improved estimator.

With the speech and noise spectral power given, MMSE estimators of the clean speech spectral Discrete Fourier Transform (DFT) coefficients can be derived, as

$$\begin{aligned}\widehat{S}_k = P\Big(\mathcal{H}_{1,k}\big|\, Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}\Big) \cdot \mathrm{E}\Big\{S_k|Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}, \mathcal{H}_{1,k}\Big\} \\ + P\Big(\mathcal{H}_{0,k}\big|\, Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}\Big) \cdot \mathrm{E}\Big\{S_k|Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}, \mathcal{H}_{0,k}\Big\}\,, \quad (1.2)\end{aligned}$$

where $P\Big(\mathcal{H}_{1,k}|\,Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}\Big)$ is the *a posteriori* Speech Presence Probability (SPP). As in speech absence the optimal clean speech estimate is zero, (1.2) can be written as [McAulay and Malpass, 1980, Ephraim and Malah, 1984, Malah *et al*, 1999]

$$\begin{aligned}\widehat{S}_k &= P\Big(\mathcal{H}_{1,k}|\,Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}\Big)\ \mathrm{E}\Big\{S_k|Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}, \mathcal{H}_{1,k}\Big\} \\ &= P\Big(\mathcal{H}_{1,k}|\,Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}\Big)\ G_{\mathcal{H}_{1,k},k}\, Y_k \\ &= G_k\, Y_k\,. \end{aligned} \tag{1.3}$$

The basic speech enhancement framework is depicted in Figure 1.1.

In this thesis we treat the estimation of the spectral gain function $G_{\mathcal{H}_{1,k},k}$ and the *a posteriori* SPP $P\Big(\mathcal{H}_{1,k}|\,Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}\Big)$ separately. Thus, in (1.3) we set $P\Big(\mathcal{H}_{1,k}|\,Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}\Big) = 1$ for all time-frequency points when we discuss the spectral gain function, while we set $G_{\mathcal{H}_{1,k},k} = 1$ for all time-frequency points when we discuss the estimation of the *a posteriori* SPP.

Assuming that the speech and noise spectral coefficients are complex Gaussian distributed, the MMSE estimator for the clean speech spectral coefficients is given by the well known Wiener Filter,

$$G_{\mathcal{H}_{1,k},k} = \mathrm{E}\Big\{S_k|Y_k, \sigma^2_{\mathrm{S},k}, \sigma^2_{\mathrm{N},k}, \mathcal{H}_{1,k}\Big\}/Y_k = \frac{\xi_k}{1+\xi_k}\,. \tag{1.4}$$

Other results than the Wiener filter may result for three reasons. First, we may search for functions of the clean speech DFT coefficients, *e.g.* the spectral amplitudes [Ephraim and Malah, 1984] or functions of the spectral amplitudes [Ephraim and Malah, 1985, You *et al*, 2005, Loizou, 2005]. Secondly, we may change the optimality criterion, *e.g.* from MMSE to Maximum *A Posteriori* (MAP) [Wolfe and Godsill, 2001]. Thirdly, the assumption on the distribution of the spectral coefficients may be changed, *e.g.* to be super-Gaussian instead of Gaussian [Martin, 2002, Martin, 2005, Lotter and Vary, 2005, Hendriks and Martin, 2007, Erkelens *et al*, 2007b]. The shape of the distribution of clean speech spectral coefficients is discussed *e.g.* in [Martin, 2002] and [Gerkmann and Martin, 2010b]. Instead of making an assumption on the distribution of speech spectral coefficients, data driven approaches can be used to derive optimal spectral gain functions [Porter and Boll, 1984, Erkelens *et al*, 2007a].

For resynthesis, each segment l of the enhanced spectrum $\widehat{S}_k(l)$ is transformed into the time domain using the Inverse Discrete Fourier Transform (IDFT), as

$$\tilde{s}_n(l) = \frac{1}{N}\sum_{k=0}^{N-1} \widehat{S}_k(l)\ \mathrm{e}^{\mathrm{j}2\pi kn/N}\,. \tag{1.5}$$

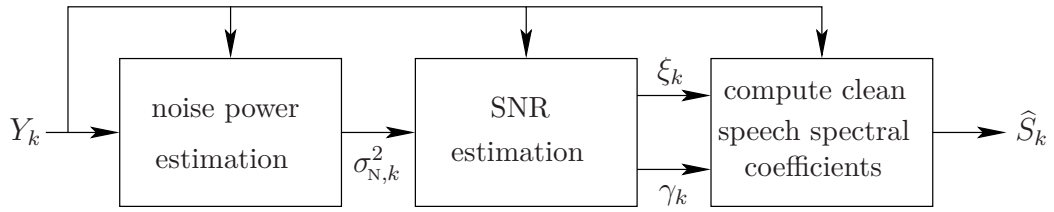


Figure 1.1: Basic speech enhancement framework.

Then, the continuous time domain signal is obtained from the segmented time domain signal $\tilde{s}_n(l)$ using the overlap-add method

$$\hat{s}(\tau) = \sum_{l\in\mathbb{Z}} \tilde{w}_{\tau-lL}\, \tilde{s}_{\tau-lL}(l)\,, \tag{1.6}$$

where the synthesis window $\tilde{w}_n$ and the segmented time domain signal $\tilde{s}_n(l)$ are assumed to be zero for $n < 0$ and $n > N-1$. For the evaluation of the speech enhancement algorithms presented in this thesis we use Hann windows w_n with a length of 32 ms and 50% overlap for the spectral analysis in (1.1) and rectangular synthesis windows $\tilde{w}_n = 1$ in (1.6). Different choices for analysis synthesis window pairs are discussed *e.g.* in [Vary and Martin, 2006, Section 11.3.5], [Mauler and Martin, 2007], and [Mauler and Martin, 2010].

1.2 Drawbacks of DFT-based speech enhancement

DFT based speech enhancement allows for a high frequency resolution that enables the suppression of noise even between the spectral harmonics of voiced sounds. However, a drawback of DFT based speech enhancement algorithms is that they may yield unnatural sounding structured residual noise, often referred to as *musical noise*, which is one of the most annoying artifacts of speech enhancement algorithms [Dreiseitel and Schmidt, 2006]. Musical noise occurs, if the spectral noise power estimate is locally underestimated. Then, *e.g.* in a noise-only signal frame single Fourier coefficients are not attenuated while all other coefficients are attenuated. The residual isolated spectral peaks in the processed spectrum correspond to sinusoids in the time domain and are perceived as tonal artifacts of one frame duration. Especially when the speech enhancement algorithms operate in nonstationary noise environments, unnatural sounding residual noise remains a challenge. On the other hand, if the estimated noise spectral

power is too large, speech distortions occur. A simple way to make musical noise inaudible and to reduce speech distortions is to apply a lower limit $G_{\min}$ on the spectral gain function, the so called *spectral floor*, as

$$\widetilde{G}_k = \max\{G_k, G_{\min}\} \ . \tag{1.7}$$

However, this also reduces the amount of noise reduction.

In this thesis, we aim at lowering the spectral floor $G_{\min}$ without introducing musical noise or speech distortions.

1.3 Related work

If certain frequency bands are known to be very strongly disturbed or even missing, a Wiener Filter cannot recover the clean speech. In such a case it may be preferable to infer the missing or strongly disturbed speech from codebooks as proposed by Rosca, Gerkmann, and Balcan [Rosca *et al*, 2006], or by using techniques from artificial bandwidth extension [Esch *et al*, 2010, Jax and Vary, 2003, Larsen and Aarts, 2004].

Isolated spectral peaks in the processed spectrum occur particularly as many DFT based speech enhancement algorithms work in each frequency bin independently. Examples for DFT based algorithms that work independently in each frequency bin are the spectral noise power estimator using Martin's minimum statistics [Martin, 2001], Ephraim and Malah's *a priori* SNR using the decision-directed approach [Ephraim and Malah, 1984], spectral gain functions such as the Wiener filter, or the iterative SPP estimator of [Malah *et al*, 1999]. The advantage of the frequency independent approach is that it can be applied to signals with arbitrary spectral shape. However, if *a priori* information about the spectral shape of the wanted signal is available, this information can be used to make the estimator more robust. Thus, some algorithms exploit the correlation of neighboring time-frequency points by smoothing spectral quantities in the time-frequency domain, *e.g.* the *a priori* SNR estimator proposed by [Fingscheidt *et al*, 2005], the smoothing of spectral gain functions proposed by [Esch and Vary, 2009], [Brandt and Bitzer, 2009], and [Gustafsson *et al*, 2001], or the SPP estimator according to [Cohen and Berdugo, 2001]. The approaches in [Srinivasan *et al*, 2006] and [Srinivasan *et al*, 2007] go further and store the representative spectral shapes of speech and noise in codebooks. However, for low input SNRs the codebook lookup becomes increasingly difficult and is prone to errors. Furthermore, the approach is rather demanding in terms of memory and computational complexity.

In this thesis, we use the cepstrum to eliminate outliers of spectral quantities which results in less musical noise. The cepstrum, which is given by the inverse Fourier transform of the logarithm of the magnitude squared spectrum, is defined in Chapter 2. One

of the first proposals to process the spectrum of speech by modifying cepstral coefficients is the nulling of the higher cepstral coefficients for formant estimation [Schafer and Rabiner, 1969]. This approach eliminates the spectral fine structure and is thus sometimes referred to as *cepstral smoothing* [Markel and Gray, 1976, Section 7.3]. The approach is based on the assumption that the transfer function of the vocal tract is represented by the lower cepstral coefficients, while the excitation of the vocal tract is represented by the higher cepstral coefficients. The resulting spectrum after nulling the higher cepstral coefficients is therefore assumed to represent the transfer function of the vocal tract.

More recently Stoica and Sandgren have proposed to null certain cepstral coefficients with a magnitude lower than a certain threshold [Stoica and Sandgren, 2006, Stoica and Sandgren, 2007] to reduce the variance of the spectral quantity. As this implies that a cepstral coefficient has zero mean if its magnitude is below the threshold, the threshold has to be carefully chosen and results in a trade-off between speech distortions and outlier suppression. In [Stoica and Sandgren, 2006, Stoica and Sandgren, 2007] the threshold is globally determined based on the standard deviation of cepstral coefficients. Thus, the approach exploits no *a priori* knowledge about which cepstral coefficients are likely to represent the speech spectral structure. As a consequence, for spectral outliers with a distinct spectral structure, such as babble bursts, the threshold is likely to be locally chosen too small, as a global increase of the threshold would also result in speech distortions.

1.4 Structure of this thesis

This thesis is organized as follows. In Chapter 2 we present the definition of the cepstrum and introduce the concept of cepstral smoothing. As a smoothing of spectral quantities in the cepstral domain results in a bias in the frequency domain, we derive a bias compensation based on a thorough analysis of the first and second order statistics of the logarithmic periodogram and cepstral coefficients. For cepstral smoothing, an estimate of the fundamental period of voiced speech sounds is required. We show that under certain assumptions a peak search in the cepstral coefficients that represent the spectral fine structure is the optimal fundamental period estimation in the maximum likelihood sense. If multiple microphones are present, we show that the optimal estimator searches for a peak in the sum of the microphone cepstra. The results are further improved by tracking the fundamental period over time.

In Chapter 3 we show how a temporal smoothing of the cepstrum may be used to improve the performance of speech enhancement algorithms. We first propose to temporally smooth the cepstral representation of the spectral gain function, and then propose to use temporal cepstrum smoothing for *a priori* SNR estimation. While the first method

is a very flexible technique that can be used for all speech enhancement algorithms that estimate the clean speech spectral coefficients via a multiplicative gain function, the latter technique increases the performance further, as it is applied in an earlier step of the speech enhancement framework.

In Chapter 4 we propose to instantaneously replace coefficients of the cepstral representation of the speech and noise spectral estimates by the corresponding cepstral coefficients of the noisy periodogram. Due to the instantaneous nature of this approach, in contrast to Temporal Cepstrum Smoothing (TCS), the representation of the residual noise is not smeared over time.

In Chapter 5 we discuss *a posteriori* Speech Presence Probability (SPP) estimation. In contrast to state-of-the-art estimators, we argue that for SPP estimation the *a priori* SPP and the *a priori* SNR should not be adapted but represent true *a priori* knowledge. We propose to smooth the *a posteriori* SNR in the frequency domain or, preferably, in the cepstral domain. The resulting *a posteriori* SPP estimators are shown to yield a better trade-off between speech distortions and noise leakage than state-of-the-art estimators.

In Chapter 6 conclusions from the results of this thesis are drawn.

In Appendix A we present mathematical derivations whose results are used in Chapter 2, while in Appendix B we present the derivation of the Maximum Likelihood (ML) *a priori* SNR estimator used in Chapter 3.

Chapter 2

Properties of the Cepstrum

In this thesis, we propose to use the cepstrum to represent the spectral characteristics of speech. The cepstrum of a positive, symmetric, real valued spectral quantity Φ_k of the speech enhancement algorithm is given by the inverse discrete Fourier transform of the natural logarithm of the spectrum

$$\phi_q = 1/N \sum_{k=0}^{N-1} \log(\Phi_k)\, \mathrm{e}^{\mathrm{j}2\pi kq/N}\,, \tag{2.1}$$

where $q \in \{0, \ldots, N-1\}$ is the cepstral index, often referred to as *quefrency* index [Bogert *et al*, 1963], and $\log(\cdot)$ is the natural logarithm. In the relevant literature, the cepstrum is sometimes defined as a function of the *magnitude* of spectral coefficients, *e.g.* $\Phi_k = |Y_k|$ [Vary and Martin, 2006], but also defined as the *squared* magnitude of spectral coefficients, *e.g.* $\Phi_k = |Y_k|^2$ [Ephraim and Rahim, 1999]. While the first definition results from the even part of the complex cepstrum defined by [Oppenheim and Schafer, 1975, Chapter 10], [Deller *et al*, 1993, Chapter 6], the latter is more in line with the introduction of the cepstrum by Bogert, Healy, and Tukey [Bogert *et al*, 1963]. See also [Oppenheim and Schafer, 2004] for a review on the history of the cepstrum. Due to the log-function, the square results only in a scaling of the cepstral coefficient by a factor of two, while the principle behavior of the cepstral coefficients remains the same. However, a scaling of the cepstrum also scales the mean and standard deviation of cepstral coefficients which are derived in Section 2.3. In this thesis, we define the cepstrum as in (2.1), where Φ_k is a *squared* quantity. Thus, Φ_k may represent spectral quantities like the noisy periodogram $|Y_k|^2$, the speech spectral power $\sigma^2_{\mathrm{s},k}$, or functions of squared quantities such as the *a posteriori* Signal-to-Noise Ratio (SNR) γ_k, or the gain function of a Wiener filter.

Note that as Φ_k is real-valued, ϕ_q is symmetric with respect to $q = N/2$. Therefore, in the following only the part $q \in \{0, \ldots, N/2\}$ is discussed. The lower cepstral coefficients $q \in \{0, \ldots, q_{\mathrm{low}}\}$ with, preferably, $q_{\mathrm{low}} \ll N/2$ represent the spectral envelope of Φ_k. To characterize all resonances of the vocal tract, assuming an allpole speech model, $q_{\mathrm{low}}/f_{\mathrm{s}}$

should be at least 1 ms [Markel and Gray, 1976, Section 6.5.3], where f_s is the sampling rate. In practice we choose q_{low}/f_s in the range of 1 ms − 4 ms. For speech signals, the spectral envelope is determined by the transfer function of the vocal tract. The higher cepstral coefficients $q_{low} < q < N/2$ represent the fine-structure of Φ_k which, for speech signals, is caused by the excitation of the vocal tract. For voiced speech, the excitation is mainly represented by a dominant peak at $q_0 = f_s/f_0$, with f_0 the fundamental frequency. The fundamental period is also represented by multiples of the fundamental period peak, the so-called *rahmonics*, rq_0, with r the index of the cepstral rahmonic. However, the energy of the rahmonics decays quickly with increasing r. The fundamental frequency can be found by a maximum search in $q \in \{q_{low}, \ldots, N/2\}$ as proposed by [Noll, 1967]. In Section 2.4 we show that, under some assumptions, this maximum search is optimal in the maximum likelihood sense. With the found fundamental frequency, in the cepstral domain voiced speech can be represented by the set

$$\mathbb{Q} = \{0, \ldots, q_{low}, \mathbb{Q}_{pitch}\} \ , \tag{2.2}$$

where $\mathbb{Q}_{pitch}$ contains the fundamental period peak and its rahmonics rq_0, with $r = \{1, \ldots, R+1\}$, where R is the considered number of rahmonics. In this thesis often only the fundamental period peak is considered, *i.e.* $R = 0$. The remaining coefficients are given by the set

$$\overline{\mathbb{Q}} = \{\{q_{low}+1, \ldots, N/2\} \setminus \mathbb{Q}_{pitch}\} \ , \tag{2.3}$$

The cepstrum is well suited for speech processing algorithms for the following reasons: while speech is mainly represented by the set $\mathbb{Q}$, non-speech like spectral structures, like spectrally narrow babble bursts, are represented mainly by the remaining cepstral coefficients $\overline{\mathbb{Q}}$. Thus, in the cepstrum speech-like and non speech-like spectral structures can be selectively treated. Furthermore, as the cepstrum is a fixed transform, the cepstral coefficients are easy to interpret. For instance the peak q_0 in the upper cepstrum can be directly related to the fundamental period T_0, as $T_0 = q_0/f_s$. Finally, the computational cost of the cepstral analysis is moderate, and dominated by the discrete Fourier transform which can be efficiently implemented via a real-valued fast Fourier transform [Cooley and Tukey, 1965, Sorensen *et al*, 1987]. In Section 3.2.3 we show how the computational complexity can be further reduced.

2.1 The cepstrum of clean speech

In Figure 2.1 the spectrum of a clean speech signal and its cepstral representation is given. Here, the logarithm of the magnitude of the cepstrum is shown. It can be seen that speech is mainly represented by some lower cepstral coefficients, a fundamental period peak and multiples of that peak, the so called rahmonics.

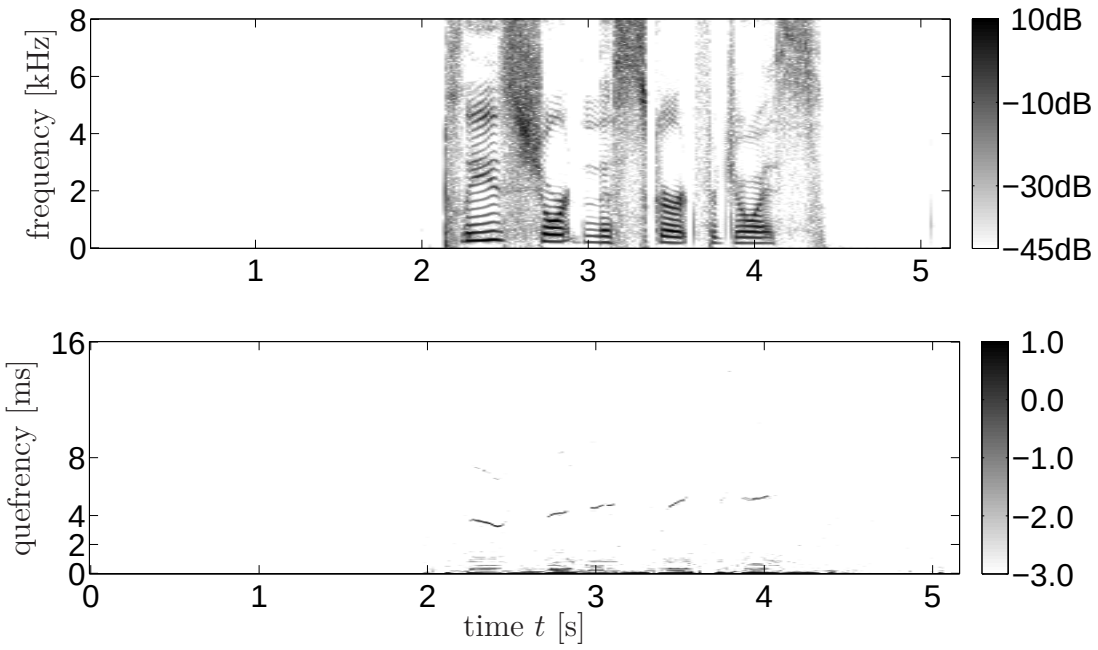


Figure 2.1: Clean spectrogram (top) and its cepstrum (bottom)

We will now show that the upper cepstral coefficients, except for the fundamental period peak, are not important to reconstruct high quality speech. To achieve this, we set the cepstral coefficients $q \in \overline{\mathbb{Q}}$ to zero. We use the Keele database [Meyer, Accessed 2006] which contains speech samples, the corresponding voiced/unvoiced information and the speech fundamental period. The speech samples are downsampled to a sampling rate of $f_\mathrm{s} = 16\,\mathrm{kHz}$. We now assess the speech quality dependent on q_low. One very common measure to compare the speech quality of processed and clean speech is the cepstral distance [Markel and Gray, 1976, Section 10.2.2], [Quackenbush *et al*, 1988, Section 2.2.7]. This measure is based on the difference between the lower cepstral coefficients of the clean and processed speech, which means that a modification of the upper cepstral coefficients is not accounted by this measure. Therefore, the cepstral distance is not well suited for this experiment. Instead, we use the Perceptual Evaluation of Speech Quality (PESQ) Mean Opinion Score (MOS) [ITU-T, 2001]. The PESQ MOS is not directly related to cepstral coefficients but measures the spectral distance between frequency bands that resemble the spectral resolution of the auditory system. In the experiment we use different settings for voiced and unvoiced speech sounds, *i.e.* we set $q_\mathrm{low} = q_\mathrm{low,v}$ for voiced sounds and $q_\mathrm{low} = q_\mathrm{low,uv}$ for unvoiced sounds. The results of the experiment are given in Figure 2.2. It may be seen, that only few cepstral coefficients are needed to obtain a large PESQ MOS. Further, as compared to unvoiced sounds, for voiced sounds more cepstral coefficients are needed to obtain a large overall PESQ MOS.

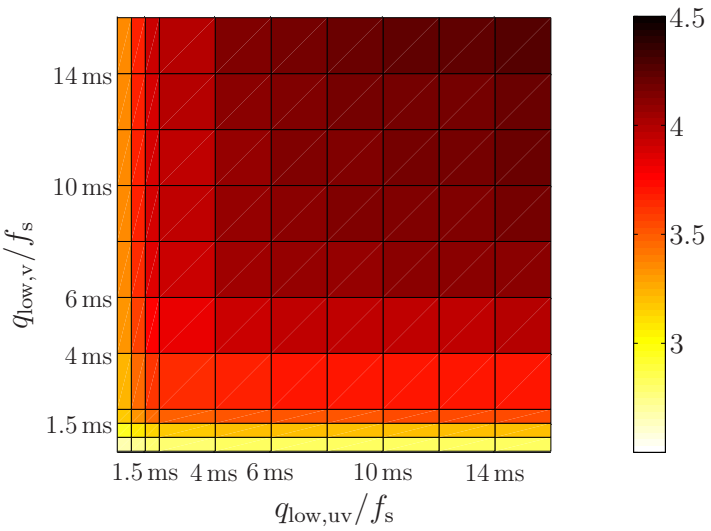


Figure 2.2: PESQ MOS for continuous speech from 5 male and 5 female speakers [Meyer, Accessed 2006]. For voiced sounds, cepstral coefficients above $q_{\mathrm{low,v}}$ are set to zero, excluding the cepstral coefficients representing the fundamental period $q \in \mathbb{Q}_{\mathrm{pitch}}$. For unvoiced sounds, cepstral coefficients above $q_{\mathrm{low,uv}}$ are set to zero. The sampling rate is $f_{\mathrm{s}} = 16\,\mathrm{kHz}$. PESQ yields values between 1 and 4.5, where 4.5 indicates the best speech quality.

2.2 Cepstral smoothing for speech enhancement without artifacts

Musical noise is caused by non-speech-like spectral outliers that alter the fine structure of a spectral quantity. In this thesis three methods for cepstral smoothing are analyzed. The first technique is to apply a selective Temporal Cepstrum Smoothing (TCS). Secondly, we may null non-speech related cepstral coefficients $q \in \overline{\mathbb{Q}}$ (Cepstral Nulling (CN)). The third technique is to instantaneously replace the non-speech related cepstral coefficients $q \in \overline{\mathbb{Q}}$ of a spectral quantity Φ_k by the corresponding coefficients of the noisy speech and is discussed in Chapter 4. In this section we will briefly introduce the concepts of TCS and CN.

The idea of TCS is that cepstral outliers of short duration can be reduced by smoothing the cepstrum over time. Due to the properties of the cepstrum, a selective smoothing of speech-like spectral structures and non speech-like spectral structures is possible. To minimize spectral distortions of the speech signal, only little smoothing is applied to the speech related cepstral coefficients $\mathbb{Q}$. On the other hand a strong smoothing can be applied to the remaining cepstral coefficients for an efficient reduction of non speech-like spectral outliers. The cepstral quantity ϕ_q, defined in (2.1), is recursively smoothed over

time, as

$$\bar{\phi}_q(l) = \alpha_q(l)\,\bar{\phi}_q(l-1) + (1-\alpha_q(l))\phi_q(l)\,, \tag{2.4}$$

where the smoothing factor α_q is chosen rather close to zero for the speech related cepstral coefficients $q \in \mathbb{Q}$ and rather close to one for the remaining coefficients $q \in \overline{\mathbb{Q}}$. The exact determination of $\mathbb{Q}$ and α_q is discussed in Chapter 3.

In CN certain cepstral coefficients are set to zero, as

$$\bar{\phi}_q = b_q \phi_q\,, \tag{2.5}$$

where the indicator function can be defined as

$$b_q = \begin{cases} 1 & \text{, for } q \in \mathbb{Q} \\ 0 & \text{, else.} \end{cases} \tag{2.6}$$

Alternatively, $b_q(l)$ can be defined to be zero if $\phi_q(l)$ is below a certain threshold, as proposed by [Stoica and Sandgren, 2007].

After the cepstral smoothing, $\bar{\phi}_q$ is transformed to the spectral domain to achieve the smoothed spectral quantity $\bar{\Phi}_k$, as

$$\bar{\Phi}_k = \mathcal{B} \cdot \exp\left(\sum_{q=0}^{N-1} \bar{\phi}_q\, \mathrm{e}^{-\mathrm{j}2\pi kq/N}\right)\,. \tag{2.7}$$

Due to the nonlinear logarithmic compression of the cepstral transform (2.1) an unbiased smoothing in the cepstrum domain leads to a bias in the spectral domain. This bias is compensated using the bias correction factor $\mathcal{B}$. The bias correction factor $\mathcal{B}$ is dependent on the distribution of the spectral quantity Φ_k. A bias correction for χ-distributed spectral amplitudes is discussed in Section 2.3, while a general bias correction for the smoothing of spectral gain functions is proposed in Section 3.1.1. In [Mauler *et al*, 2008] it is shown how cepstral smoothing can be interpreted as a smoothing on the entire time-frequency plane. Further in [Mauler *et al*, 2008] we also present a bias compensation for χ^2-distributed spectral quantities. However, as opposed to the bias compensation in Section 2.3, the bias correction proposed in [Mauler *et al*, 2008] holds only for spectrally uncorrelated spectral coefficients, is computationally rather expensive, and can only be applied to TCS.

The TCS algorithm is summarized in Algorithm 1, while the algorithm for CN is summarized in Algorithm 2.

Algorithm 1 Temporal Cepstrum Smoothing (TCS) of a spectral quantity Φ_k, such as a spectral gain function, a speech power estimate, or the *a posteriori* SNR.

1: **for all** signal segments l **do**
2: Compute the cepstrum of Φ_k (2.1)
$$\phi_q = 1/N \sum_{k=0}^{N-1} \log(\Phi_k)\ \mathrm{e}^{\mathrm{j}2\pi kq/N}\,.$$
3: Choose smoothing factor α_q to be rather close to zero for the speech related cepstral coefficients $q \in \mathbb{Q}$ and rather close to one for the remaining coefficients $q \in \overline{\mathbb{Q}}$.
4: Apply selective Temporal Cepstrum Smoothing (TCS) (2.4)
$$\bar{\phi}_q(l) = \alpha_q(l)\,\bar{\phi}_q(l-1) + (1-\alpha_q(l))\phi_q(l)\,.$$
5: Compute the bias correction $\mathcal{B}$.
6: Transform back into the frequency domain (2.7)
$$\bar{\Phi}_k = \mathcal{B} \cdot \exp\left(\sum_{q=0}^{N-1} \bar{\phi}_q\, \mathrm{e}^{-\mathrm{j}2\pi kq/N}\right)\,.$$
7: **end for**

Algorithm 2 Cepstral Nulling Cepstral Nulling (CN) of a spectral quantity Φ_k, such as a spectral gain function, a speech power estimate, or the *a posteriori* SNR.

1: **for all** signal segments l **do**
2: Compute the cepstrum of Φ_k (2.1)
$$\phi_q = 1/N \sum_{k=0}^{N-1} \log(\Phi_k)\ \mathrm{e}^{\mathrm{j}2\pi kq/N}\,.$$
3: Choose b_q as
$$b_q = \begin{cases} 1 & \text{, for } q \in \mathbb{Q} \\ 0 & \text{, else.} \end{cases}$$
4: Apply selective Cepstral Nulling (CN) (2.5)
$$\bar{\phi}_q = b_q \phi_q\,.$$
5: Compute the bias correction $\mathcal{B}$.
6: Transform back into the frequency domain (2.7)
$$\bar{\Phi}_k = \mathcal{B} \cdot \exp\left(\sum_{q=0}^{N-1} \bar{\phi}_q\, \mathrm{e}^{-\mathrm{j}2\pi kq/N}\right)\,.$$
7: **end for**

2.3 Statistical properties of cepstral coefficients and χ-distributed spectral amplitudes before and after cepstral smoothing

In this section we show that if χ-distributed spectral amplitudes are smoothed in the cepstral domain, the resulting smoothed spectral amplitudes are also approximately χ-distributed but with more degrees of freedom and less signal power. Further, we provide new insights into the statistics of the cepstral coefficients derived from χ-distributed spectral amplitudes using tapered spectral analysis windows. We derive explicit expressions for the variance and covariance of correlated χ-distributed spectral amplitudes and the resulting cepstral coefficients, parameterized by the degrees of freedom. Finally, we derive the signal power bias $\mathcal{B}$ that arises when spectral amplitudes are smoothed by reducing their variance in the cepstral domain by means of a cepstral smoothing via TCS or CN and develop a power bias compensation method. The proposed bias correction results in a simple scaling of the spectral amplitudes and is fixed for a fixed set of cepstral smoothing parameters. As the determination of the bias correction factor is computationally inexpensive, it can be computed on a segment-by-segment basis if the smoothing parameters change. The results of this section are partly presented in [Gerkmann and Martin, 2009].

In many applications of statistical signal processing, a variance reduction of spectral quantities derived from time domain signals, such as the periodogram, is required [Martin, 2001, Gerkmann *et al*, 2008b]. The χ^2-distribution of a spectral quantity P is given as

$$p(P) = \frac{1}{\Gamma(\mu)} \left(\frac{\mu}{\sigma^2}\right)^{\mu} P^{\mu-1} \exp\!\left(-\frac{\mu}{\sigma^2} P\right) , \tag{2.8}$$

with shape parameter μ, mean $\mathrm{E}\{P\} = \sigma^2$, variance $\mathrm{var}\{P\} = \sigma^4/\mu$, and the complete gamma function $\Gamma(\cdot)$ [Gradshteyn and Ryzhik, 2000, (8.31)]. 2μ is also known as the *degrees of freedom* [Vary and Martin, 2006]. Mean and variance can be derived using [Gradshteyn and Ryzhik, 2000, (3.381.4)]. For $\mu = 1$, it is well known that a smoothing of P over time and/or frequency results in an approximately χ^2-distributed random variable with the same mean and an increase in the degrees of freedom that goes along with the decreased variance [Martin and Lotter, 2001, Martin, 2006]. In Appendix A.1 we show that the χ^2-distribution holds exactly for a moving average smoothing of independent periodogram bins P and arbitrary μ. Then, for an unbiased smoothing, the shape parameter after smoothing $\bar{\mu}$ can be easily obtained if the mean before smoothing and the amount of variance reduction is known, as

$$\bar{\mu} = \sigma^4 / \mathrm{var}\left\{\bar{P}\right\} , \tag{2.9}$$

where smoothed quantities are marked by a bar. A drawback of smoothing in the frequency domain is that the temporal and/or frequency resolution is reduced. In speech processing this may not be desired as the temporal smoothing smears speech onsets and frequency smoothing reduces the resolution of speech harmonics. This drawback of frequency domain smoothing is overcome by cepstral smoothing techniques. However, the application of an unbiased smoothing process in the cepstral domain leads to a bias in the spectral domain: cepstral smoothing does not only change the variance of a χ^2-distributed spectral random variable P, but also its mean $\mathrm{E}\{P\} = \sigma^2 \neq \mathrm{E}\left\{\bar{P}\right\} = \bar{\sigma}^2$. For practical applications, the fact that cepstral smoothing results in a bias in the frequency domain is most critical. If $P = |S|^2$ is the periodogram of a complex-valued zero-mean variable S for instance, changing the mean of the periodogram $\mathrm{E}\{|S|^2\}$ changes the signal power of S. As this is usually an undesired side-effect of cepstral smoothing, a framework to compensate for the bias in signal power is needed. However, after cepstral smoothing, all three variables in (2.9) are unknown. We neither know the shape parameter after smoothing $\bar{\mu}$, nor the amount of variance reduction a cepstral smoothing applies to spectral coefficients, nor the biased mean of the spectral coefficients after smoothing $\bar{\sigma}^2$. In this section we show that bias, the variance, and the shape parameter in (2.9) can still be determined based on a statistical analysis of the log-periodogram and cepstral coefficients. The presented results are based on the observation that the distribution of spectral amplitudes after cepstral smoothing can be well approximated by a χ-distribution. Then we show that the variance of the cepstral coefficients is directly related to the shape parameter. Thus, for a given amount of variance reduction in the cepstral domain, we can determine the shape parameter after smoothing. The bias can then be determined from the shape parameters before and after smoothing.

We first discuss the statistical properties of the log-periodogram and of cepstral coefficients in Section 2.3.1 for several spectral analysis windows. In Section 2.3.2 we show how the shape parameter after cepstral smoothing can be determined and how the signal power bias can be compensated. This procedure is summarized in Algorithm 4. In Section 2.3.3 we discuss the mean of the cepstral coefficients. In Section 2.3.4 we apply the proposed bias compensation in a practical scenario.

2.3.1 Statistical properties of the logarithmic periodogram and cepstral coefficients before cepstral smoothing

It is well known that for a Gaussian time domain signal $s(\tau)$, the spectral coefficients S_k obtained similar to (1.1) are complex Gaussian distributed and the spectral amplitudes $|S_k|$ are χ-distributed with two degrees of freedom ($\mu = 1$) for $k \in \{\{1, \ldots, N\} \setminus N/2\}$, and with one degree of freedom ($\mu = 1/2$) at $k \in \{0, N/2\}$. The χ-distribution is given

by

$$p(|S_k|) = \frac{2}{\Gamma(\mu)} \left(\frac{\mu}{\sigma^2_{\mathrm{S},k}}\right)^{\mu} |S_k|^{2\mu-1} \exp\left(-\frac{\mu}{\sigma^2_{\mathrm{S},k}}|S_k|^2\right) , \tag{2.10}$$

with the variance $\sigma^2_{\mathrm{S},k} = \mathrm{E}\{|S_k|^2\}$. The distribution of the periodogram $P_k = |S_k|^2$ is then found to be the χ^2-distribution [Papoulis and Pillai, 2002],

$$p(P_k) = \frac{1}{\Gamma(\mu)} \left(\frac{\mu}{\sigma^2_{\mathrm{S},k}}\right)^{\mu} P_k^{\mu-1} \exp\left(-\frac{\mu}{\sigma^2_{\mathrm{S},k}}P_k\right) . \tag{2.11}$$

The χ^2-distribution is frequently also referred to as the Gamma distribution [Andrianakis and White, 2009]. χ and χ^2-distribution can also be comprised to a *generalized Gamma* distribution [Erkelens *et al*, 2007b]. For $\mu = 1$ the χ-distribution is identical to the Rayleigh distribution, while the χ^2-distribution is identical to the exponential distribution. As stated above, a shape parameter of $\mu = 1$ results for a Gaussian distributed time domain signal $s(\tau)$. Even if the time domain signal $s(\tau)$ is not Gaussian distributed, the complex spectral coefficients are asymptotically Gaussian distributed for large N [Brillinger, 1981]. However, for segment sizes used in common speech processing frameworks, it can be shown that the complex spectral coefficients of speech signals are super-Gaussian distributed [Martin, 2002, Martin, 2005] and thus exhibit a larger kurtosis as compared to Gaussian distributed spectral coefficients. The kurtosis of the χ-distribution (2.10) can be shown to increase with a decreasing μ [Breithaupt, 2008, (C.1)]. In fact, choosing $\mu < 1$ in (2.10) may yield a better fit to the distribution of speech spectral amplitudes than a Rayleigh distribution ($\mu = 1$) [Andrianakis and White, 2006, Breithaupt *et al*, 2008b, Andrianakis and White, 2009].

In this thesis, we derive expressions for arbitrary values of μ that thus hold for complex Gaussian distributed spectral coefficients S_k ($\mu = 1$), complex super-Gaussian distributed spectral coefficients for $\mu < 1$ [Breithaupt and Martin, 2010] and complex spectral coefficients that exhibit a slightly sub-Gaussian distribution for $\mu > 1$. In a practical scenario, μ should be chosen so that (2.10) fits the empirical distribution of the spectral amplitudes of the considered signal. However, we show in this thesis that μ can also be estimated from the empirical variance of cepstral coefficients (cf. Algorithm 3).

To compute the variance of the cepstral coefficients we first derive the variance of the log-periodogram,

$$\mathrm{var}\{\log(P_k)\} = \mathrm{E}\left\{(\log(P_k))^2\right\} - (\mathrm{E}\{\log(P_k)\})^2 . \tag{2.12}$$

With (2.11) and [Gradshteyn and Ryzhik, 2000, (4.352.1)], the expected value of the log-periodogram of a χ^2-distributed P_k can be derived as

$$\mathrm{E}\{\log P_k\} = \psi(\mu) - \log(\mu) + \log\left(\sigma^2_{\mathrm{S},k}\right) , \tag{2.13}$$

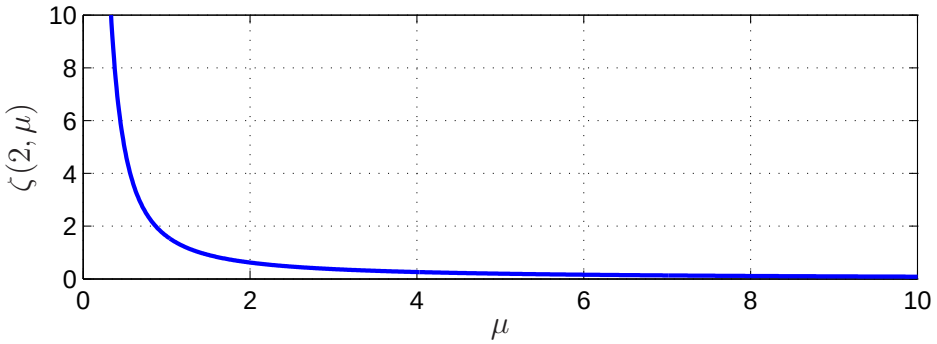


Figure 2.3: Riemann's zeta-function $\zeta(2,\mu)$ [Gradshteyn and Ryzhik, 2000, (9.521.1)]

where $\psi(\cdot)$ is the psi-function [Gradshteyn and Ryzhik, 2000, (8.360)]. The first term on the right hand side of (2.12) can be derived using [Gradshteyn and Ryzhik, 2000, (4.358.2)]

$$\mathrm{E}\Big\{(\log P_k)^2\Big\} = \Big(\psi(\mu) - \log(\mu) + \log\Big(\sigma^2_{\mathrm{S},k}\Big)\Big)^2 + \zeta(2,\mu)\ , \tag{2.14}$$

where

$$\zeta(2,\mu) = \sum_{n=0}^{\infty} \frac{1}{(\mu+n)^2}\,. \tag{2.15}$$

is Riemann's zeta-function [Gradshteyn and Ryzhik, 2000, (9.521.1)], depicted in Figure 2.3.

With (2.12), (2.13), and (2.14) the variance of the log-periodogram κ_0 results in

$$\kappa_0 = \mathrm{var}\{\log P_k\} = \zeta(2,\mu)\ . \tag{2.16}$$

This is a generalization of the results in [Ephraim and Rahim, 1999], where the variance of the log-periodogram was derived for the special case $\mu = 1$.

The cepstrum is obtained as given in (2.1) with $\Phi_k = P_k = |S_k|^2$. As shown in Appendix A.2, the covariance of the cepstral coefficients can be obtained by taking a two dimensional discrete Fourier transform of the covariance of the log-periodogram as

$$\mathrm{cov}\{\phi_{q_1}, \phi_{q_2}\} = \frac{1}{N^2} \sum_{k_2=0}^{N-1} \sum_{k_1=0}^{N-1} \mathrm{cov}\{\log(P_{k_1})\,, \log(P_{k_2})\}\, e^{\mathrm{j}\frac{2\pi}{N} q_1 k_1} e^{-\mathrm{j}\frac{2\pi}{N} q_2 k_2}\,, \tag{2.17}$$

where $k_1, k_2 \in \{0, \ldots, N-1\}$ are frequency indices, and $q_1, q_2 \in \{0, \cdots, N/2\}$ are quefrency indices. For large N, we may neglect the fact that at $k \in \{0, N/2\}$ the variance

$\mathrm{var}\{\log P_k\} = \zeta(2, \frac{\mu}{2})$ is larger than for $k \in \{\{1, \ldots, N\} \setminus N/2\}$ where $\mathrm{var}\{\log P_k\} = \zeta(2, \mu) = \kappa_0$. If the frequency bins are uncorrelated, *i.e.* $\mathrm{cov}\{\log P_{k_1}, \log P_{k_2}\} = 0$ for $k_1 \neq k_2$, the covariance of the cepstral coefficients results in

$$\mathrm{cov}\{\phi_{q_1}, \phi_{q_2}\} = \begin{cases} \frac{1}{N}\kappa_0 & , q_1 = q_2, q_1 \in \left\{1, ..., \frac{N}{2} - 1\right\} \\ \frac{2}{N}\kappa_0 & , q_1 = q_2, q_1 \in \left\{0, \frac{N}{2}\right\} \\ 0 & , q_1 \neq q_2 \end{cases}, \tag{2.18}$$

with κ_0 defined in (2.16). Note that a tapered spectral analysis window w_n in (1.1) results in a correlation of adjacent frequency bins. Since in (2.18) uncorrelated frequency bins are assumed, this result holds only for rectangular spectral analysis windows. Tapered spectral analysis windows and correlated spectral coefficients are treated in the following paragraph.

Correlated spectral coefficients and tapered spectral analysis windows

While in [Ephraim and Rahim, 1999] and (2.18) only rectangular spectral analysis windows w_n were considered for the spectral analysis in (1.1), we now discuss the statistics of the log-periodogram and cepstral coefficients for correlated spectral coefficients, where the spectral correlation results *e.g.* from tapered spectral analysis windows as used in many speech processing algorithms.

While for uncorrelated spectral coefficients we have μ degrees of freedom for $k \in \{0, N/2\}$ and 2μ degrees of freedom for $k \in \{\{1, \ldots, N\} \setminus N/2\}$, the correlation introduced by a tapered spectral analysis window results in a reduction of the degrees of freedom, and thus a higher variance for the log-periodogram bins adjacent to $k = 0$ and $k = N/2$. As for large N this hardly affects the cepstral coefficients, the effect is insignificant here. A derivation of the log-spectral variances is given by [Gray, Jr, 1974] for the special case $\mu = 1$ and different spectral analysis windows w_n.

However, the squared correlation coefficient of the frequency coefficient S_k and its mth neighbor S_{k+m}

$$\rho_m^2 = \frac{\left|\mathrm{E}\left\{S_k S_{k+m}^*\right\}\right|^2}{\mathrm{E}\{|S_k|^2\}\mathrm{E}\{|S_{k+m}|^2\}} \tag{2.19}$$

greatly affects the variance of cepstral coefficients.

The resulting covariance of the logarithm of two periodogram bins

$$\kappa_m = \mathrm{cov}\{\log(P_k), \log(P_{k+m})\}$$

is derived below. As in general the spectral covariance κ_m introduced by tapered spectral analysis windows rapidly decreases with increasing m, we assume that $\kappa_m = 0$ for $m > M$ and $M \ll N/2+1$. Further, as in (2.18), for large N, we may neglect the fact that for $k \in \{0, N/2\}$ the variance of the log-periodogram is larger than κ_0, as we have less degrees of freedom than for $k \notin \{0, N/2\}$. Then, as shown in Appendix A.3, the covariance of cepstral coefficients ϕ_{q_1} and ϕ_{q_2} for correlated data results in

$$\operatorname{cov}\{\phi_{q_1}, \phi_{q_2}\} \approx \begin{cases} \frac{1}{N}\left(\kappa_0 + 2\sum_{m=1}^{M} \kappa_m \cos\left(m\frac{2\pi}{N}q_1\right)\right) & , q_1 = q_2, q_1 \in \left\{1, ..., \frac{N}{2}-1\right\} \\ \frac{2}{N}\left(\kappa_0 + 2\sum_{m=1}^{M} \kappa_m \cos\left(m\frac{2\pi}{N}q_1\right)\right) & , q_1 = q_2, q_1 \in \left\{0, \frac{N}{2}\right\} \\ 0 & , q_1 \neq q_2 \end{cases} \tag{2.20}$$

where M denotes the number of non-zero covariance values κ_m. From (2.20) it follows that cepstral coefficients are approximately uncorrelated, even if log-periodogram bins are correlated. The cepstral variance is given as the diagonal of the covariance matrix, as

$$\operatorname{var}\{\phi_q\} \approx \begin{cases} \frac{2}{N}\left(\zeta(2,\mu) + 2\sum_{m=1}^{M} \kappa_m \cos\left(m\frac{2\pi}{N}q\right)\right) & , q \in \left\{0, \frac{N}{2}\right\} \\ \frac{1}{N}\left(\zeta(2,\mu) + 2\sum_{m=1}^{M} \kappa_m \cos\left(m\frac{2\pi}{N}q\right)\right) & , \text{else} \end{cases} . \tag{2.21}$$

To derive the covariance κ_m of two log-periodogram bins, we extend the χ^2-distribution (2.11) to the bivariate χ^2-distribution of two correlated periodogram bins $P_k = |S_k|^2$ and $P_{k+m} = |S_{k+m}|^2$ with the squared correlation coefficient ρ_m as given in (2.19). This distribution can be found *e.g.* in [Joarder, 2009, Theorem 2.1], as

$$p(P_k, P_{k+m}) = \frac{P_k^{\mu-1} P_{k+m}^{\mu-1}}{2^{2\mu+1}\sqrt{\pi}\,\Gamma(\mu)\,(1-\rho_m^2)^{\mu}} \exp\left(-\frac{P_k + P_{k+m}}{2(1-\rho_m^2)}\right) \sum_{n=0}^{\infty} \left(1 + (-1)^n\right) \left(\frac{\rho_m}{1-\rho_m^2}\right)^n \frac{\Gamma\left(\frac{n+1}{2}\right)}{n!\,\Gamma\left(\frac{n}{2}+\mu\right)} P_k^{\frac{n}{2}} P_{k+m}^{\frac{n}{2}} . \tag{2.22}$$

Note that the infinite sum in (2.22) can also be expressed in terms of the hypergeometric function [Nadarajah, 2009]. With (2.22), [Gradshteyn and Ryzhik, 2000, (4.352.1)] and [Gradshteyn and Ryzhik, 2000, (3.381.4)] we find

$$\kappa_m = \operatorname{cov}\{\log(P_k), \log(P_{k+m})\} \tag{2.23}$$

$$= \mathrm{E}\{\log(P_k)\log(P_{k+m})\} - \mathrm{E}\{\log(P_k)\}\mathrm{E}\{\log(P_{k+m})\} \tag{2.24}$$

$$= \sum_{n=0}^{\infty} A_{n,\mu,\rho_m} \left(B_{n,\mu,\rho_m}\right)^2 - \left(\sum_{n=0}^{\infty} A_{n,\mu,\rho_m} B_{n,\mu,\rho_m}\right)^2 , \tag{2.25}$$

where

$$A_{n,\mu,\rho_m} = \frac{(1-\rho_m^2)^\mu}{2\sqrt{\pi}\,\Gamma(\mu)}\,(1+(-1)^n)\,2^n\rho_m^n\frac{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{n}{2}+\mu\right)}{n!}\,, \tag{2.26}$$

$$B_{n,\mu,\rho_m} = \psi\left(\mu+\frac{n}{2}\right)+\log\left(2\left(1-\rho_m^2\right)\right), \tag{2.27}$$

and ρ_m as defined in (2.19). This is a generalization of the results in [Ephraim and Roberts, 2005, (6)] and [Ephraim and Roberts, 2005, (20)] where the covariance is given for the special cases $\mu = 1$ and $\mu = 1/2$, respectively. With (2.25), the covariance κ_m of log-periodogram bins, and thus the covariance of cepstral coefficients (2.20), can be determined.

From above derivations we see that the covariance of cepstral coefficients depends only on the shape parameter μ of χ-distributed spectral amplitudes and the correlation between spectral coefficients ρ_m. Specifically, the covariance of the cepstral coefficients is independent of the signal power, the spectral shape, and the segment index l. In Appendix A.4 we show that for a Hann window and $\sigma^2_{\mathrm{S},k-1} \approx \sigma^2_{\mathrm{S},k} \approx \sigma^2_{\mathrm{S},k+1}$, the normalized correlation results in $\rho_1 = 2/3$ and $\rho_2 = 1/6$. Hence, for a Hann window and $\mu = 1$ we have $\kappa_1 = 0.507$ and $\kappa_2 = 0.028$.

The cepstral variance for $\mu = 1$ and the rectangular window ($\kappa_m = 0, m \in \{1, ..., M\}$) or the Hann window ($\kappa_1 = 0.507, \kappa_2 = 0.028, \kappa_m = 0, m \in \{3, ..., M\}$) are compared in Figure 2.4, where we also show empirical data. It is obvious that (2.21) provides an excellent fit for both the rectangular and the Hann window. As the additional cosine-terms in (2.20) and (2.21) have zero mean with respect to q, the mean of the cepstral variance for arbitrary spectral correlation equals the cepstral variance for a rectangular window and is thus independent of the chosen analysis window w_n. Thus, for the sum over quefrency we have

$$\sum_{q=0}^{N/2}\nu_q \mathrm{var}\{\phi_q\} = \zeta(2,\mu)\ , \tag{2.28}$$

with

$$\nu_q = \begin{cases} 1/2 & , q \in \{0, N/2\} \\ 2 & , \text{else} \end{cases}\ . \tag{2.29}$$

The coefficients ν_q account for the symmetry of the cepstrum and the different variances at the DC and Nyquist bin in (2.21). In this way the cosine terms in (2.21) cancel out and the modified summed variance of the cepstral coefficients are related to the shape parameter μ via Riemann's zeta-function.

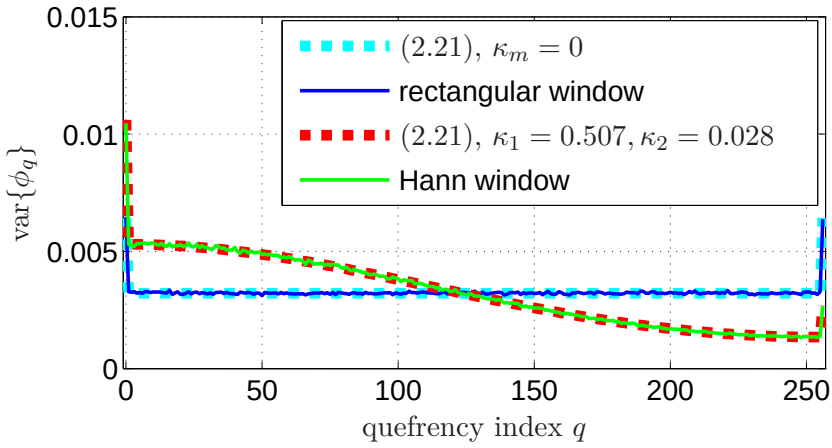


Figure 2.4: The cepstral variance for a pink Gaussian time-domain signal analyzed with a non-overlapping rectangular analysis window w_n in (2) and a Hann window with half-overlapping frames. The empirical variances are compared to the theoretical results in (2.21) with $\kappa_m = 0, m \in \{1, ..., M\}$ for the rectangular window and $\kappa_1 = 0.507, \kappa_2 = 0.028, \kappa_m = 0, m \in \{3, ..., M\}$ for the Hann window. The sampling rate is $f_\mathrm{s} = 16\,\mathrm{kHz}$ and $N = 512$.

2.3.2 Statistical properties after cepstral smoothing

In this section, we approximate the distribution of spectral amplitudes after cepstral smoothing by the parametric χ-distribution. From experimental results in Section 2.3.4 it will be seen that this approximation is valid. From (2.28) and Figure 2.3 we see that a reduction of the cepstral variance via cepstral smoothing increases the parameter μ of the χ-distribution. Then, due to (2.13), changing μ also changes the spectral power $\sigma^2_{\mathrm{s},k}$. Hence, a variance reduction in the cepstral domain results in a bias in the spectral power that can now be accounted for. In the following, we denote parameters after cepstral smoothing by a bar. We will discuss cepstral smoothing via CN and TCS separately.

As described in Section 2.2, in CN a set of cepstral coefficients is set to zero. Then, the summed variance after cepstral smoothing can be related to the shape parameter after smoothing $\bar{\mu}$, as

$$\zeta(2, \bar{\mu}) = \sum_{q=0}^{N/2} \nu_q \mathrm{var}\{\phi_q\}\, b_q\,, \tag{2.30}$$

where $\bar{\phi}_q$ are the cepstral coefficients after cepstral smoothing, the indicator function $b_q \in \{0, 1\}$ sets certain cepstral coefficients to zero, and ν_q is defined as in (2.29).

For TCS, the cepstral coefficients are recursively smoothed over time with a quefrency dependent smoothing factor α_q

$$\bar{\phi}_q(l) = \alpha_q \, \bar{\phi}_q(l-1) + (1-\alpha_q) \, \phi_q(l) \, . \tag{2.31}$$

With the variance after recursive smoothing derived in Appendix A.5.1, the variance after cepstral smoothing can be related to the shape parameter after smoothing $\bar{\mu}$, as

$$\zeta(2, \bar{\mu}) = \sum_{q=0}^{N/2} \nu_q \mathrm{var}\{\phi_q\} \, \frac{1-\alpha_q}{1+\alpha_q} \, . \tag{2.32}$$

The derivation in Appendix A.5.1 holds for uncorrelated successive signal segments which is valid for nonoverlapping rectangular spectral analysis windows and is also well fulfilled for half overlapping Hann windows. For higher signal segment correlation, the summed variance after cepstral smoothing can be measured offline for a fixed set of recursive smoothing constants α_q. For a given μ of the spectral amplitudes before cepstral smoothing, the cepstral variance can be determined via (2.21) and thus the summed cepstral variance after cepstral smoothing via (2.30) or (2.32). In a practical application, the relation between $\bar{\mu}$ and $\zeta(2, \bar{\mu})$ can be stored in a table such that the summed cepstral variance can be directly related to the shape parameter after cepstral smoothing $\bar{\mu}$.

The spectral power bias $\sigma^2_{\mathrm{S},k}/\bar{\sigma}^2_{\mathrm{S},k}$ can then be determined using (2.13), as

$$\begin{aligned} &\log\left(\sigma^2_{\mathrm{S},k}/\bar{\sigma}^2_{\mathrm{S},k}\right) \\ &= \mathrm{E}\Big\{\log\left(|S_k|^2\right)\Big\} - \psi(\mu) + \log(\mu) - \Big(\mathrm{E}\Big\{\overline{\log\left(|S_k|^2\right)}\Big\} - \psi(\bar{\mu}) + \log(\bar{\mu})\Big) \, . \end{aligned} \tag{2.33}$$

The cepstral transformation consists of a nonlinear logarithmic compression and an Inverse Discrete Fourier Transform (IDFT). As the IDFT is a linear operation, an unbiased smoothing in the cepstral domain remains unbiased in the logarithmic domain. Therefore the expectation of the logarithmic periodogram stays unchanged before and after cepstral smoothing, *i.e.* $\mathrm{E}\{\log(|S_k|^2)\} = \mathrm{E}\Big\{\overline{\log(|S_k|^2)}\Big\}$. We thus obtain the frequency independent factor

$$\mathcal{B} = \sigma^2_{\mathrm{S},k}/\bar{\sigma}^2_{\mathrm{S},k} = \frac{\mu}{\bar{\mu}} \exp(\psi(\bar{\mu}) - \psi(\mu)) \tag{2.34}$$

that is applied when computing the inverse cepstral transform as in (2.7), where $\bar{\Phi}_k$ represents the periodogram after cepstral smoothing, *i.e.* the squared spectral amplitudes. Note that the bias correction $\mathcal{B}$ depends only on μ and $\bar{\mu}$. For a fixed set of smoothing parameters α_q or b_q the bias correction $\mathcal{B}$ is thus fixed and independent of

the segment index l. We obtain unbiased cepstrally-smoothed spectral amplitudes with reduced cepstral variance, as

$$|\bar{S}_k(l)| = \sqrt{\mathcal{B} \exp\left(\sum_{q=0}^{N-1} \bar{\phi}_q(l)\ \mathrm{e}^{-\mathrm{j}2\pi kq/N}\right)}\ , \tag{2.35}$$

which are approximately χ-distributed (2.10) with shape parameter $\bar{\mu}$. The algorithm for computing unbiased signal power estimates after cepstral smoothing is summarized in Algorithm 4, while Algorithm 3 summarizes how to obtain the statistical properties before cepstral smoothing that are required for Algorithm 4.

Algorithm 3 Determination of second order statistics before cepstral smoothing

1: If unknown, determine the shape parameter μ using an empirical estimation of $\mathrm{var}\{\phi_q\}$ from representative data and (2.28):

$$\zeta(2,\mu) = \sum_{q=0}^{N/2} \nu_q \mathrm{var}\{\phi_q\}\ ,$$

with ν_q defined in (2.29).

2: Determine the correlation of neighboring log-periodogram bins κ_m via (2.25):

$$\kappa_m = \sum_{n=0}^{\infty} A_{n,\mu,\rho_m} \left(B_{n,\mu,\rho_m}\right)^2 - \left(\sum_{n=0}^{\infty} A_{n,\mu,\rho_m} B_{n,\mu,\rho_m}\right)^2$$

with A, B, ρ_m defined in (2.26), (2.27), and (2.19).

3: Determine the cepstral variance before cepstral smoothing (2.21):

$$\mathrm{var}\{\phi_q\} = \begin{cases} \frac{2}{N}\left(\zeta(2,\mu) + 2\sum_{m=1}^{M} \kappa_m \cos\left(m\frac{2\pi}{N}q\right)\right) & , q \in \left\{0, \frac{N}{2}\right\} \\ \frac{1}{N}\left(\zeta(2,\mu) + 2\sum_{m=1}^{M} \kappa_m \cos\left(m\frac{2\pi}{N}q\right)\right) & , \text{else.} \end{cases}$$

2.3.3 Mean of the cepstrum

In this section we derive the mean of the cepstral coefficients. We generalize the results of [Ephraim and Rahim, 1999] and [Stoica and Sandgren, 2006, Stoica and Sandgren, 2007], where $\mu = 1$ is assumed. Due to the linearity of the inverse discrete Fourier transform IDFT$\{\cdot\}$ and (2.13), the mean value of the cepstral coefficients defined by

Algorithm 4 Bias compensation for Temporal Cepstrum Smoothing (TCS) and Cepstral Nulling (CN)

1: Determine the cepstral variance before smoothing using Algorithm 3.
2: **for all** signal segments l **do**
3: **if** smoothing parameters b_q or α_q have changed **then**
4: Determine the shape parameter after cepstral smoothing $\bar{\mu}$,

- in the case of CN (2.30):

$$\zeta(2,\bar{\mu}) = \sum_{q=0}^{N/2} \nu_q \mathrm{var}\{\phi_q\}\, b_q,$$

- in the case of TCS (2.32):

$$\zeta(2,\bar{\mu}) = \sum_{q=0}^{N/2} \nu_q \mathrm{var}\{\phi_q\}\, \frac{1-\alpha_q}{1+\alpha_q}.$$

5: Compute signal power bias (2.34):

$$\mathcal{B}(\bar{\mu}) = \sigma^2_{\mathrm{s},k}/\bar{\sigma}^2_{\mathrm{s},k} = \frac{\mu}{\bar{\mu}} \exp(\psi(\bar{\mu}) - \psi(\mu)) \ .$$

6: **end if**
7: Apply bias correction when computing the inverse cepstral transform (2.7):

$$\bar{\Phi}_k(l) = \mathcal{B}(\bar{\mu}) \exp\left(\sum_{q=0}^{N-1} \bar{\phi}_q(l)\ \mathrm{e}^{-\mathrm{j}2\pi kq/N} \right) \ .$$

8: **end for**

In a practical application, the relation between $\bar{\mu}$ and $\zeta(2,\bar{\mu})$ can be stored in a table.

(2.1) is given by

$$\begin{aligned}\mathrm{E}\{\phi_q\} =&\mathrm{IDFT}\{\mathrm{E}\{\log P_k\}\}\\ =&\mathrm{IDFT}\Big\{\log\sigma^2_{\mathrm{S},k}\Big\} - \mathrm{IDFT}\{\log\mu_k - \psi(\mu_k)\}\\ =&\mathrm{IDFT}\Big\{\log\sigma^2_{\mathrm{S},k}\Big\} - \varepsilon_q\,,\end{aligned} \tag{2.36}$$

where $\psi(\cdot)$ is the psi-function [Gradshteyn and Ryzhik, 2000, (8.360)]. Therefore, even for white signals, when $\sigma^2_{\mathrm{S},k}$ is constant over frequency, the mean of the cepstral coefficients is not zero for $q > 0$ but $-\varepsilon_q$. When

$$\mu_k = \begin{cases}\mu/2 & ,k \in \{0, N/2\}\\ \mu & ,\text{else}\end{cases}$$

the deviation ε_q results in

$$\begin{aligned}\varepsilon_q &= \mathrm{IDFT}\{\log\mu_k - \psi(\mu_k)\}\\ &= \begin{cases}\frac{N-2}{N}\left(\log\mu - \psi(\mu)\right) + \frac{2}{N}\left(\log\frac{\mu}{2} - \psi\Big(\frac{\mu}{2}\Big)\right) & ,q = 0\\ \frac{2}{N}\left(\log\frac{\mu}{2} - \psi\Big(\frac{\mu}{2}\Big)\right) - \frac{2}{N}\left(\log\mu - \psi(\mu)\right) & ,q \text{ even}\\ 0 & ,q \text{ odd}\end{cases}.\end{aligned} \tag{2.37}$$

If $\mu_k = \mu$ is constant for all k, as assumed in [Stoica and Sandgren, 2006, Stoica and Sandgren, 2007], the deviation results in

$$\varepsilon_q = \begin{cases}\log(\mu) - \psi(\mu) & ,q = 0\\ 0 & ,\text{else}\end{cases}.$$

For CN proposed by [Stoica and Sandgren, 2006] cepstral coefficients below a variance threshold are nulled, implying they have zero mean. Thus, for CN better performance can be expected when the cepstrum actually has zero mean for white signals. Such an alternative definition of the cepstrum is given by $\phi_q \leftarrow \phi_q + \varepsilon_q$. However, since typically $\varepsilon_q^2 \ll \mathrm{var}\{\phi_q\}$ for $q > 0$, the influence of the mean bias ε_q given in (2.37) is of minor importance. For TCS, as proposed in [Breithaupt *et al*, 2007, Breithaupt *et al*, 2008a], zero-mean cepstral coefficients are neither assumed nor required.

2.3.4 Experimental results

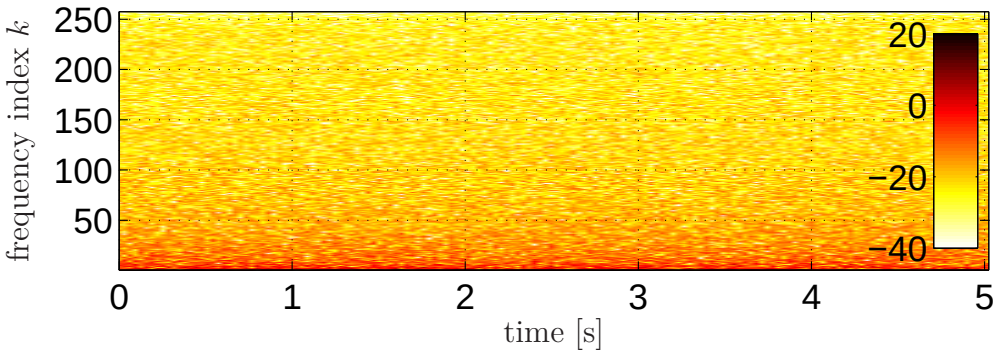
In this section we show that Algorithm 4 successfully compensates for the signal power bias introduced by cepstral smoothing. After providing results for a stationary colored signal, we also apply the bias compensation in a practical scenario, namely the *a priori* speech power estimation proposed in [Breithaupt *et al*, 2008a]. The smoothed spectral

amplitudes with bias correction are denoted by a bar and obtained as given in (2.35). When the bias compensation $\mathcal{B}$ is not applied, this results in the biased smoothed spectral amplitudes denoted by a tilde and given by

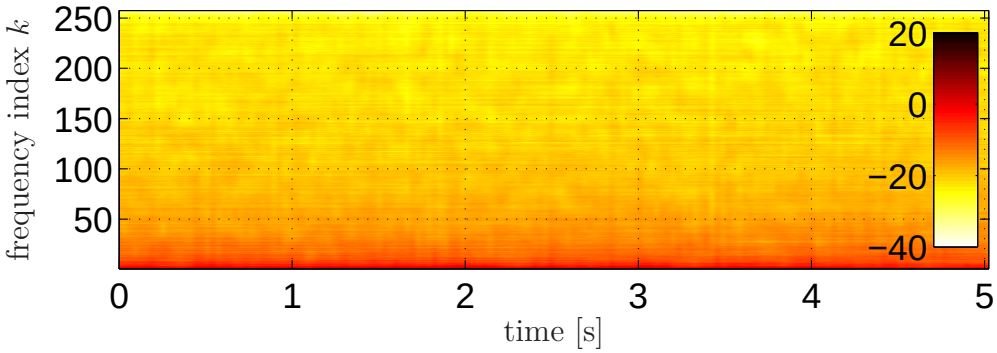
$$|\widetilde{S}_k(l)| = \sqrt{\exp\left(\sum_{q=0}^{N-1} \bar{\phi}_q(l)\ \mathrm{e}^{-\mathrm{j}2\pi kq/N}\right)}\,. \tag{2.38}$$

Stationary colored signal

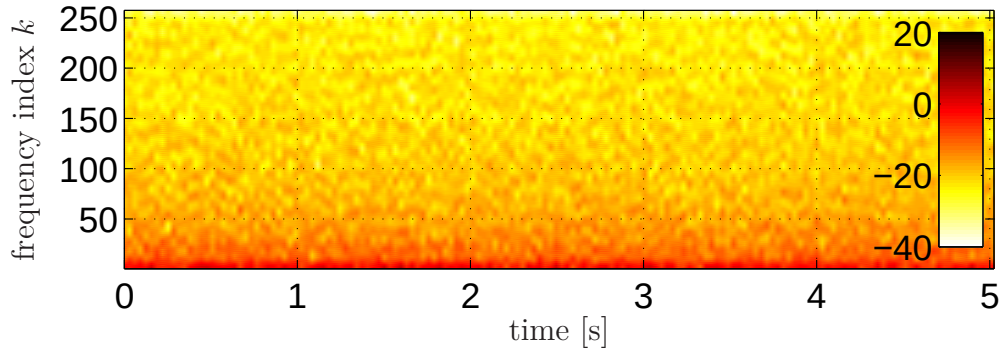
Here we apply cepstral smoothing to a stationary colored Gaussian distributed signal. The according spectrograms before and after cepstral smoothing are given in Figure 2.5. In Figure 2.6 and Figure 2.7 we present the frame energy and histograms for TCS using a rectangular and a Hann spectral analysis window in (1.1), respectively. In Figure 2.8 and Figure 2.9 we present the frame energy and histograms for CN using a rectangular and a Hann spectral analysis window in (1.1), respectively. From the presented results, we see that cepstral smoothing introduces a signal power bias, and that this bias is successfully compensated with Algorithm 4. Further, we compare the histograms of spectral amplitudes before and after cepstral smoothing with and without a bias compensation to the derived probability density functions. It may be seen that the algorithm for estimating the shape parameter after cepstral smoothing works well, as an excellent match for the histograms and the derived probability density functions may be observed. The distribution of χ-distributed spectral amplitudes after cepstral smoothing can thus be well approximated by a χ-distribution with an increased shape parameter.



(a) Spectogram of pink noise

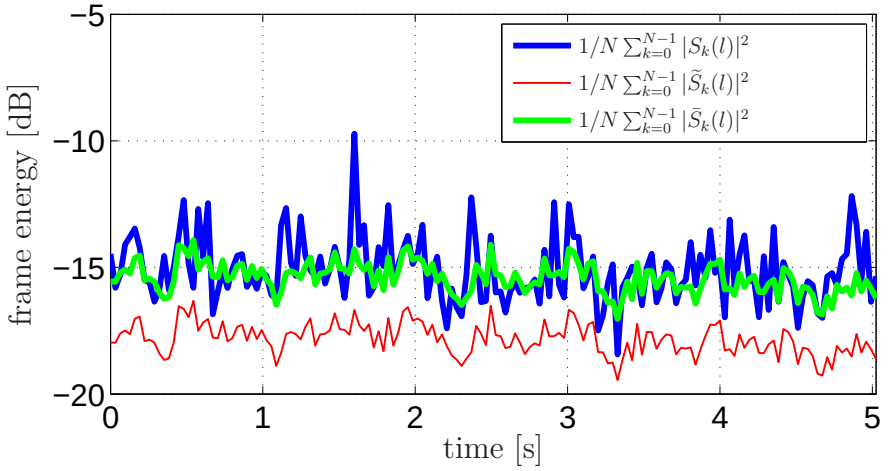


(b) Pink noise after TCS

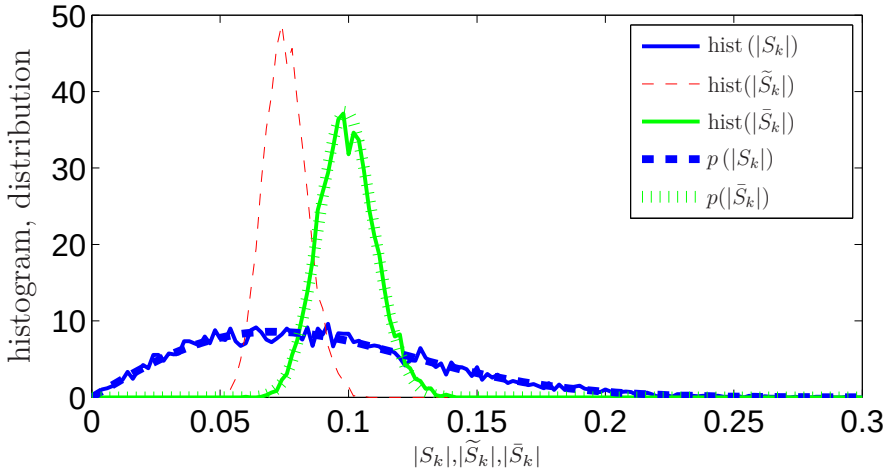


(c) Pink noise after CN

Figure 2.5: Spectrogram of Gaussian-distributed pink noise (a), after TCS (b) and after CN (c). For TCS we use the same smoothing constants as in [Breithaupt *et al*, 2008a] while for CN cepstral coefficients $q > N/16$ are set to zero. Here $N = 512$ and the sampling rate is $f_s = 16\,\text{kHz}$.

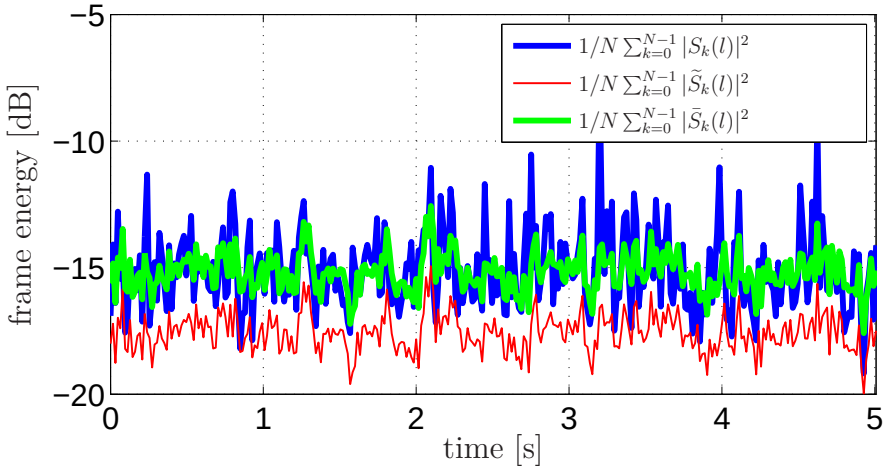


(a)

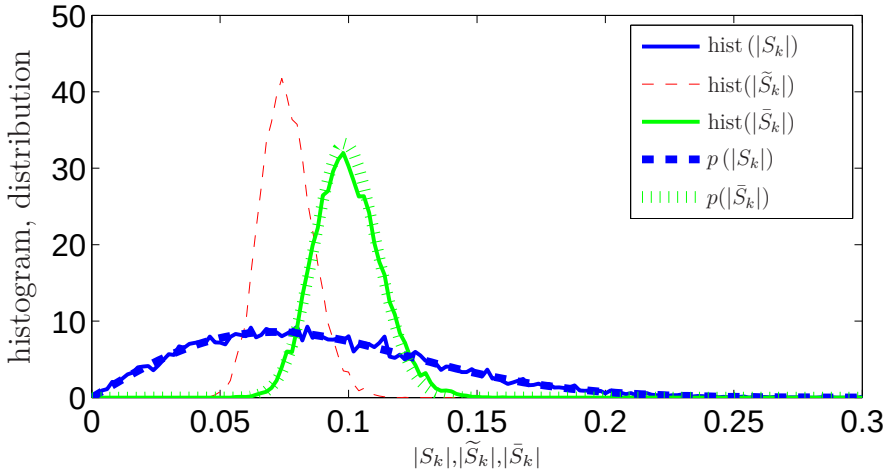


(b)

Figure 2.6: Frame energy and histograms for cepstral smoothing by TCS of a stationary pink Gaussian-distributed signal and non-overlapping rectangular spectral analysis windows w_n in (1.1). We use the same smoothing constants as in [Breithaupt *et al*, 2008a]. The spectrograms before and after processing are given in Figure 2.5(a) and Figure 2.5(b). In subplot (a) of this figure, the signal segment energies before cepstral smoothing, after cepstral smoothing, and after cepstral smoothing and bias correction are given. (b) compares the derived distributions to the histograms of the spectral amplitudes $|S_k|$ for $k = 111$ before cepstral smoothing, $|\widetilde{S}_k|$ after cepstral smoothing, and $|\bar{S}_k|$ after cepstral smoothing and bias correction.

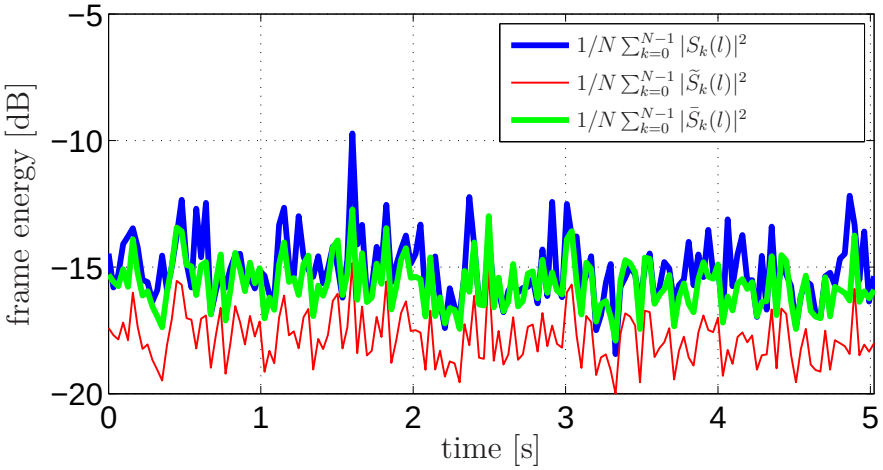


(a)

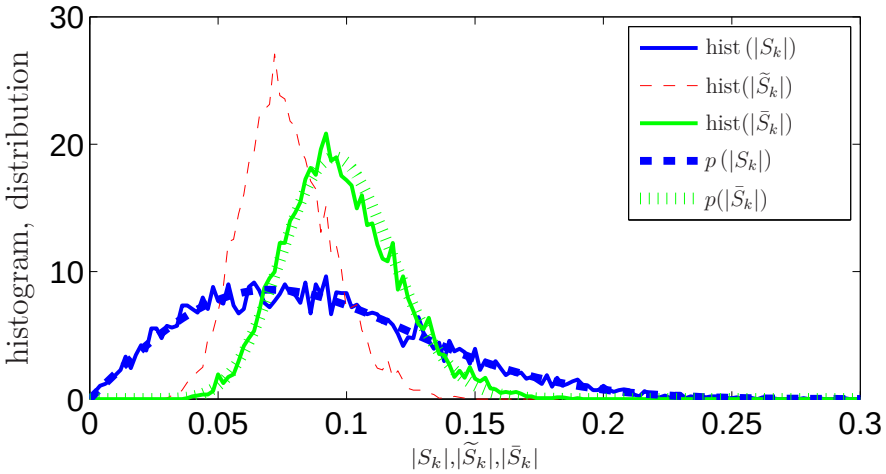


(b)

Figure 2.7: cepstral smoothing by TCS as in Figure 2.6 but with half-overlapping Hann windows w_n in (2). In (a) the signal segment energies before cepstral smoothing, after cepstral smoothing, and after cepstral smoothing and bias correction are given. (b) compares the derived distributions to the histograms of the corresponding spectral amplitudes.

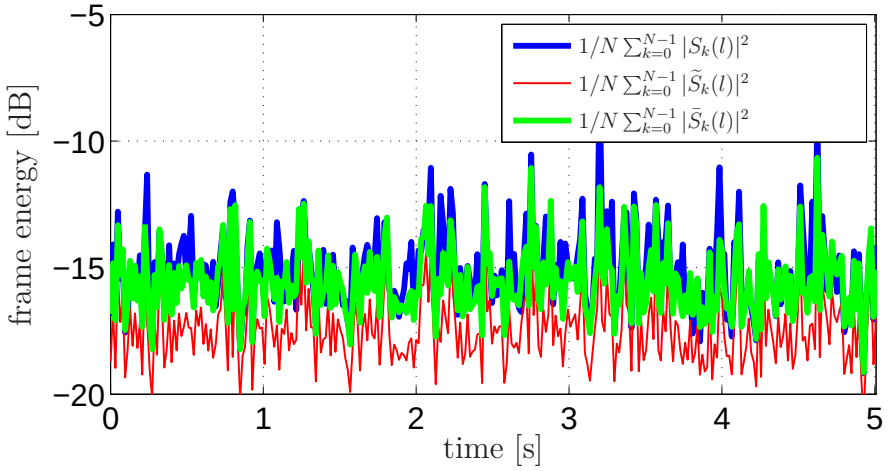


(a)

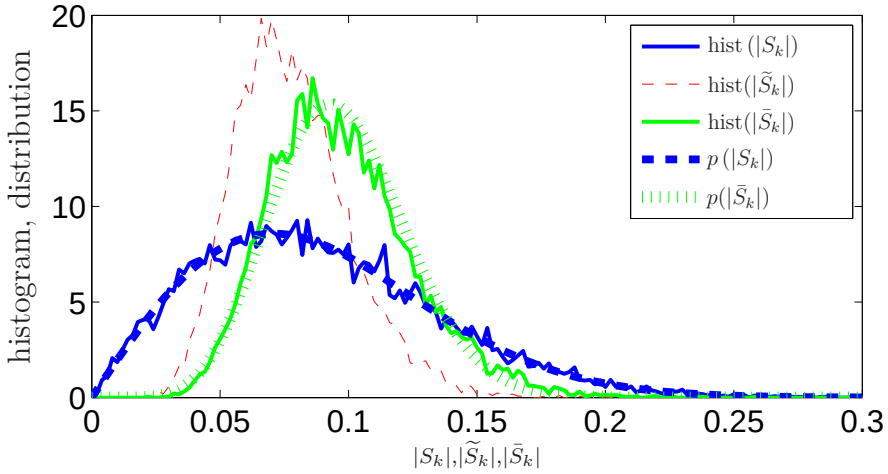


(b)

Figure 2.8: Cepstral smoothing by CN using non-overlapping rectangular spectral analysis windows w_n in (2). Cepstral coefficients $q > N/16$ are set to zero. The spectrograms before and after processing are given in Figure 2.5(a) and Figure 2.5(c). In subplot (a) of this figure, the signal segment energies before cepstral smoothing, after cepstral smoothing, and after cepstral smoothing and bias correction are given. (b) compares the derived distributions to the histograms of the corresponding spectral amplitudes.



(a)



(b)

Figure 2.9: Cepstral smoothing by CN as in Figure 2.8 but with half-overlapping Hann windows w_n in (2). In (a) the signal segment energies before cepstral smoothing, after cepstral smoothing, and after cepstral smoothing and bias correction are given. (b) compares the derived distributions to the histograms of the corresponding spectral amplitudes.

A priori speech power estimation

Now, the bias compensation method is applied in a practical scenario, namely the TCS based *a priori* clean speech power estimation algorithm for speech enhancement as proposed in [Breithaupt *et al*, 2008a]. There, a maximum likelihood estimation of the *a priori* clean speech power Φ_k is temporally smoothed in the cepstral domain via Algorithm 1 to obtain the smoothed *a priori* speech power estimation $\bar{\Phi}_k$. Without the bias correction, the *a priori* speech power estimate $\widetilde{\Phi}_k$ is biased with respect to Φ_k, as may be seen in Figure 2.10.

For the simulation we used nonoverlapping rectangular spectral analysis windows. We estimate the shape parameter before cepstral smoothing, 2μ, by measuring the average cepstral variance $\overline{\mathrm{var}\{\phi_q\}}$ and using the relation $\zeta(2,\mu) = N\overline{\mathrm{var}\{\phi_q\}}$. We thus obtain $\mu = 0.37$. We use the same smoothing procedure as proposed in [Breithaupt *et al*, 2008a]. As in [Breithaupt *et al*, 2008a] the smoothing constant α_q in (2.31) varies from signal segment to signal segment, a different bias $\mathcal{B}$ is introduced in each segment l. Note that the computational simplicity of Algorithm 4 allows for an individual computation of the signal power bias $\mathcal{B}$ in each signal segment l (steps 4-5 of Algorithm 4).

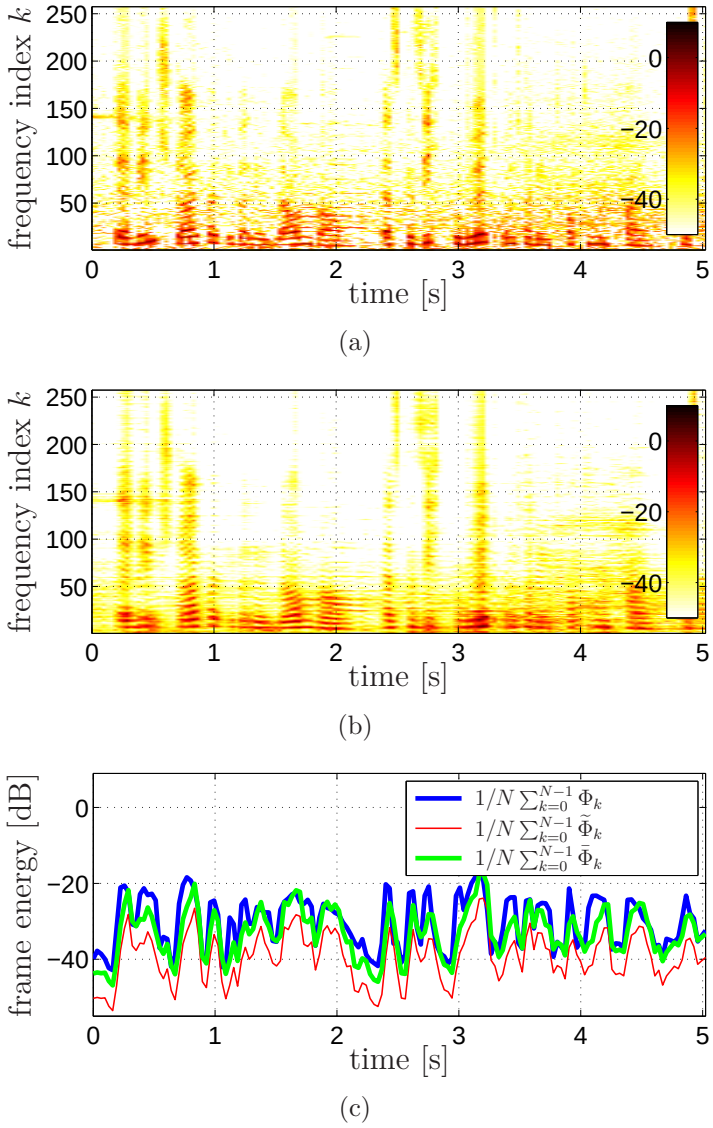


Figure 2.10: Spectrogram of the *a priori* speech power estimation Φ_k before cepstral smoothing (a) and $\bar{\Phi}_k$ after cepstral smoothing and bias compensation (b). In (c) the signal segment energies before cepstral smoothing, after cepstral smoothing, and after cepstral smoothing and bias compensation are given. The speech signal is disturbed by instationary traffic noise at a signal-to-noise ratio of 0 dB. The spectral noise power is estimated using the minimum statistics approach [Martin, 2001]. Here $N = 512$ and the sampling rate is $f_s = 16\,\text{kHz}$.

2.4 Maximum a posteriori fundamental period estimation in the cepstral domain

In this section we show that for uncorrelated spectral coefficients a maximum search in the upper cepstrum $q > q_{\text{low}}$ is the optimal cepstral domain fundamental period estimator in the Maximum Likelihood (ML) sense (Section 2.4.2). Further, we show that for multiple microphones, the ML fundamental period estimator results in a maximum search on the sum of the microphone cepstra (Section 2.4.3). Finally, we extend the ML estimator towards a Maximum *A Posteriori* (MAP) optimal fundamental period tracker (Section 2.4.4). In Section 2.4.5 we show that the proposed ML estimator outperforms a maximum search on the cepstrum of the output signal of a delay-and-sum beamformer for various input signal-to-noise ratios. The extension towards a MAP fundamental period tracker is shown to substantially increase the robustness in noisy environments. The results of this section are partly presented in [Gerkmann *et al*, 2009].

The fundamental period of voiced speech is caused by vibrations of the glottis. Its inverse, the fundamental frequency, is often simply referred to as *pitch*. As the speech fundamental period is one of the most important speech parameters, many solutions for fundamental period estimation have been proposed [Hess, 1983]. The fundamental period may be estimated for instance in the time domain using harmonic modelling [Tabrikian *et al*, 2004], the autocorrelation function [Cheveigné and Kawahara, 2002], exploiting the impulse-like characteristic of glottal excitations [Yegnanarayana and Murty, 2009], or in the cepstral domain [Noll, 1967]. Knowledge about the speech fundamental period may be exploited for instance in speech coding [Vary and Martin, 2006], and speech enhancement [Tilp, 2002, Breithaupt *et al*, 2007, Breithaupt *et al*, 2008a]. As most algorithms in this thesis require a fundamental period estimation in the cepstral domain, cepstrum domain optimal fundamental period estimators are of particular interest.

In (2.2) we define the set of speech related cepstral coefficients $\mathbb{Q} = \{0, \ldots, q_{\text{low}}, \mathbb{Q}_{\text{pitch}}\}$, where the lower cepstral coefficients $q \leq q_{\text{low}}$ represent the transfer function of the vocal tract and the set $\mathbb{Q}_{\text{pitch}}$ represents the excitation of the vocal tract for voiced sounds. With the sampling frequency f_{s}, the fundamental period T_0 of the excitation signal of voiced sounds is represented by a dominant peak in the upper cepstrum at $q_0 = T_0 f_{\text{s}}$, and multiples of that peak, the so-called *rahmonics* [Noll, 1967, Bogert *et al*, 1963] at rq_0 with $r \in \{1, 2, \ldots\}$. Thus, Noll suggests to search for the maximum peak of the squared cepstrum in the range of quefrencies that corresponds to the fundamental period [Noll, 1967].

2.4.1 Distribution of cepstral coefficients

To derive the ML fundamental period estimator, the distribution of cepstral coefficients is needed. A common assumption for the cepstral coefficients is that they are Gaussian distributed with fixed variance [Stoica and Sandgren, 2006]. As shown in Section 2.3, the variance (2.21) is dependent on the distribution of spectral coefficients and the spectral correlation. The mean of the cepstral coefficient is given by the spectral shape and a mean deviation ε_q (2.36). The mean deviation can easily be determined as given in (2.37). Thus, if ε_q is added to the cepstral coefficients, as proposed in Section 2.3.3, we can assume that the cepstral coefficients $q \in \overline{\mathbb{Q}}$ have zero mean and a variance given by (2.21). However, the influence of ε_q on the cepstral coefficients is usually rather small for $q > 0$, as then $\varepsilon_q^2 \ll \mathrm{var}\{\phi_q\}$. In [Ephraim and Rahim, 1999] it has been shown that the cepstral coefficients are asymptotically uncorrelated for large N. Thus, under the Gaussian assumption, cepstral coefficients are asymptotically independent [Hyvärinen *et al*, 2001], and their joint distribution factors into marginal distributions.

2.4.2 ML fundamental period estimator

In this section we derive a ML estimator for the fundamental period in the cepstral domain.

Because the mean of the cepstrum is zero for $q \neq rq_0$ and $q > q_{\mathrm{low}}$, the distribution of a noisy cepstral observation vector $\boldsymbol{\phi} = \left[\phi_{q_{\mathrm{low}}}, \phi_{q_{\mathrm{low}}+1}, ..., \phi_{N/2-1}\right]^T$ given the speech fundamental period index q_0 can be written as

$$\begin{aligned} p(\boldsymbol{\phi}|q_0) &= \prod_{q=q_{\mathrm{low}}+1}^{N/2-1} \frac{1}{(2\pi\sigma_q^2)^{\frac{1}{2}}} \exp\left(-\frac{(\phi_q - \mathrm{E}\{\phi_q\})^2}{2\sigma_q^2}\right) \\ &= \frac{1}{(2\pi)^{\frac{N/2-q_{\mathrm{low}}-1}{2}}} \left(\prod_{q=q_{\mathrm{low}}+1}^{N/2-1} \frac{1}{\sigma_q}\right) \exp\left(-\sum_{q=q_{\mathrm{low}}+1}^{N/2-1} \frac{\phi_q^2}{2\sigma_q^2}\right) \\ &\quad \cdot \exp\left(\sum_{r=1}^{R+1} \frac{2\phi_{rq_0}\mathrm{E}\{\phi_{rq_0}\} - (\mathrm{E}\{\phi_{rq_0}\})^2}{2\sigma_{rq_0}^2}\right). \end{aligned} \tag{2.39}$$

For simplicity we neglect the Nyquist bin $q = N/2$, as even for uncorrelated spectral coefficients it has a different variance than the coefficients $q_{\mathrm{low}} < q < N/2$ (2.18).

As the first part of (2.39) is independent of q_0, only the second exponential function has to be evaluated. As the exponential function is monotonically increasing the ML

estimator is given by

$$\begin{aligned} q_0^{\mathrm{ML}} &= \arg\max_{q_0} p(\boldsymbol{\phi}|q_0) \\ &= \arg\max_{q_0} \sum_{r=1}^{R+1} \frac{2\phi_{rq_0}\mathrm{E}\{\phi_{rq_0}\} - (\mathrm{E}\{\phi_{rq_0}\})^2}{\sigma_{rq_0}^2} \\ &= \arg\max_{q_0} \sum_{r=1}^{R+1} \frac{\mathrm{E}\{\phi_{rq_0}\}\,(2\phi_{rq_0} - \mathrm{E}\{\phi_{rq_0}\})}{\sigma_{rq_0}^2}\,. \end{aligned}$$

For uncorrelated spectral coefficients, the cepstral variance is constant for $q \notin \{0, N/2\}$ (2.18), and the ML estimator simplifies to

$$q_0^{\mathrm{ML}} = \arg\max_{q_0} \sum_{r=1}^{R+1} \mathrm{E}\{\phi_{rq_0}\}\,(2\phi_{rq_0} - \mathrm{E}\{\phi_{rq_0}\})\,.$$

Search on the squared cepstrum

As speech is highly nonstationary and hence not ergodic, the estimation of the expected value $\mathrm{E}\{\phi_{rq_0}\}$ is difficult. A simple but reasonable solution is to take the instantaneous value, as $\widehat{\mathrm{E}}\{\phi_{rq_0}\} = \phi_{rq_0}$. Then, the ML fundamental period estimation results in a peak search on the normalized squared cepstrum

$$q_0^{\mathrm{ML}} = \arg\max_{q_0} \sum_{r=1}^{R+1} \left(\phi_{rq_0}^2/\sigma_{rq_0}^2\right)\,. \tag{2.40}$$

Thus, for uncorrelated spectral coefficients and $R = 0$ a peak detection on the squared cepstrum is an optimal fundamental period estimator in the ML sense, as

$$q_0^{\mathrm{ML},R=0} = \arg\max_{q} \phi_q^2\,. \tag{2.41}$$

Note that this corresponds to the fundamental period estimator proposed in [Noll, 1967], where the cepstrum is defined equivalently to $(\phi_q N)^2$, *i.e.* the square of N times (2.1).

Search for a positive peak

Due to the symmetry of the logarithmic spectrum, the cepstral transform (2.1) results in a correlation of the log-spectrum with cosine functions. As spectral harmonics have the same fixed distance between each other and the zeroth spectral coefficient, the cepstral fundamental period peak that results from the correlation of a cosine with the spectral structure is positive as illustrated by an exemplary simulated voiced speech sound in Figure 2.11(a). If rectangular spectral analysis windows w_n are used in (1.1), the rahmonics are also positive (cf. example in Figure 2.11(b)). However, if the spectral harmonics are broader than one bin, the rahmonics may also become negative as illustrated in Figure 2.11(d). This occurs, for instance, when tapered spectral analysis windows are used that result in a smearing of spectral harmonics. The smearing results from the convolution of the the spectral harmonics with the rather broad mainlobe of the frequency response of the tapered spectral analysis window.

The *a priori* knowledge that a fundamental period peak (and for rectangular spectral analysis windows also the rahmonics) are positive, can be exploited to increase the robustness of the fundamental period peak estimator by excluding negative values from the peak search. This can be achieved by using the absolute instantaneous value for the estimate of the expected value, as $\widehat{\mathrm{E}}\{\phi_{rq_0}\} = |\phi_{rq_0}|$. The corresponding ML fundamental period estimation results in

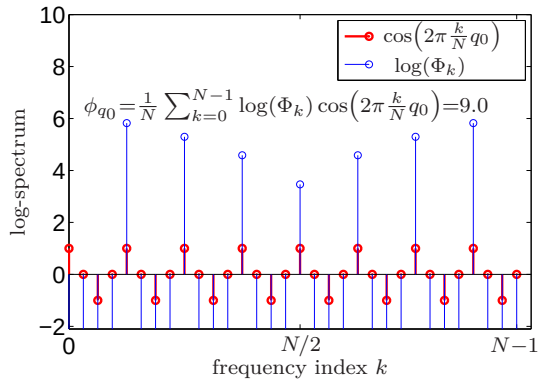
$$q_0^{\mathrm{ML}} \;=\; \arg\max_{q_0} \sum_{r=1}^{R+1} |\phi_{rq_0}| \left(2\phi_{rq_0} - |\phi_{rq_0}|\right) / \sigma_{rq_0}^2 \,. \tag{2.42}$$

As a result, negative values are penalized by a factor of three, and for $R = 0$ we obtain

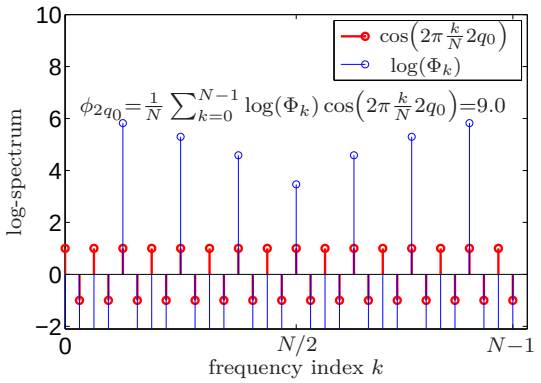
$$q_0^{\mathrm{ML},R=0} \;=\; \arg\max_{q} \begin{cases} \phi_q^2/\sigma_q^2 & ,\phi_q \geq 0 \\ -3\phi_q^2/\sigma_q^2 & ,\phi_q < 0\,. \end{cases} \tag{2.43}$$

Thus, for uncorrelated spectral coefficients and $R = 0$, the ML optimal fundamental period estimator is given by the search for a positive peak, as

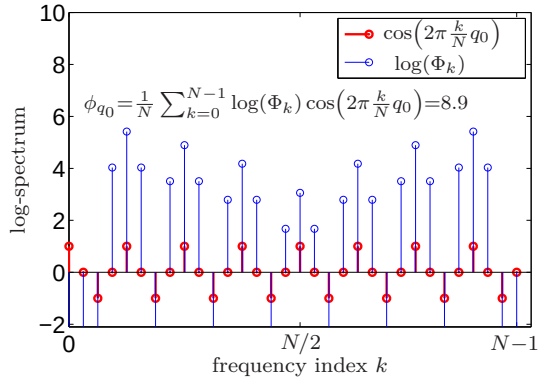
$$q_0^{\mathrm{ML},R=0} = \arg\max_{q} \phi_q \,. \tag{2.44}$$



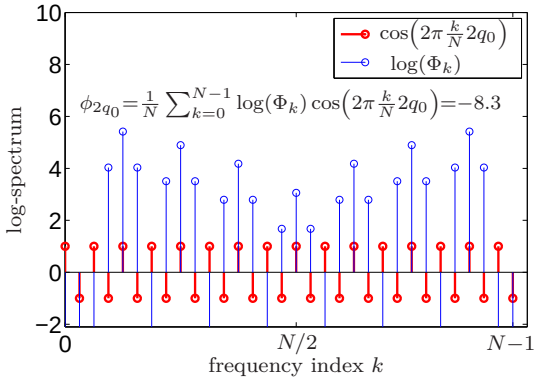
(a) For a rectangular spectral analysis window the 0th rahmonic $\phi_{q_0} = 9.0$ is positive.



(b) For a rectangular spectral analysis window the 1st rahmonic $\phi_{2q_0} = 9.0$ is positive.



(c) For a Hann spectral analysis window the 0th rahmonic $\phi_{q_0} = 8.9$ is positive.



(d) For a Hann spectral analysis window the 1st rahmonic $\phi_{2q_0} = -8.3$ is negative.

Figure 2.11: Simulation of the log-spectrum a voiced speech sound. Between the spectral harmonics, the ideal spectrum would be zero and the log spectrum would tend to minus infinity. Here, we limited the log-spectrum to be larger than -36. To compute the cepstrum, the log-spectrum is correlated with cosine functions. For a rectangular spectral analysis, in this example the zeroth and first cepstral rahmonic are equally strong (cf. figures 2.11(a) and 2.11(b)). However, if a Hann spectral analysis window is used, the rahmonics can also be negative (cf. Figure 2.11(d)).

Smoothing the cepstrum over quefrency

For a better estimate of the expected value operator, the cepstrum can be smoothed over quefrency. The smoothing can be obtained by convolving the cepstrum *e.g.* with a normalized Hamming window $w_{\mathrm{H},q}$, as

$$\check{\phi}_q = \phi_q * w_{\mathrm{H},q}\,. \tag{2.45}$$

The smoothing of the cepstrum over quefrency can also be seen as a low pass filtering of the log-spectrum. Since the power of voiced sounds is less at high frequencies, *e.g.* [Loizou, 2007, Section 4.2], the quefrency smoothing (2.45) can be expected to increase the robustness of the proposed algorithm.

Given the smoothed cepstrum (2.45), the expected value is approximated as $\widehat{\mathrm{E}}\{\phi_{rq_0}\} = |\check{\phi}_{rq_0}|$, and the ML estimator results in

$$q_0^{\mathrm{ML}} \;=\; \arg\max_{q_0} \sum_{r=1}^{R+1} |\check{\phi}_{rq_0}| \left(2\phi_{rq_0} - |\check{\phi}_{rq_0}|\right) / \sigma^2_{rq_0}\,. \tag{2.46}$$

For uncorrelated spectral coefficients, this results in

$$q_0^{\mathrm{ML}} \;=\; \arg\max_{q_0} \sum_{r=1}^{R+1} |\check{\phi}_{rq_0}| \left(2\phi_{rq_0} - |\check{\phi}_{rq_0}|\right)\,. \tag{2.47}$$

We observed that for $R = 0$ very similar results are obtained if we search for a positive peak on the normalized smoothed cepstrum, as

$$q_0^{\mathrm{ML},R=0} \;\approx\; \arg\max_{q} \check{\phi}_q\,. \tag{2.48}$$

while for $R = 1$ the performance degrades if we use $q_0^{\mathrm{ML}} \approx \arg\max_q \sum_{r=1}^{R+1} \check{\phi}_{rq_0}$ instead of (2.47).

2.4.3 Extension to multiple microphones

To extend the ML optimal solution towards the case when M microphones are present, we assume that the cepstral coefficients, *given* q_0, of the M microphones are independent. As we condition the likelihood on q_0 and consider only $q > q_{\mathrm{low}}$, this corresponds to the

assumption that the non speech related cepstral coefficients $q \in \overline{\mathbb{Q}}$ between microphones are independent. Thus, we can write

$$p(\boldsymbol{\Phi}|q_0) = \prod_{m=1}^{M} p(\boldsymbol{\phi}_m|q_0)\,, \tag{2.49}$$

with $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_M]$. For $R = 0$, and a quefrency and microphone independent cepstral variance σ_q^2, the ML estimator for multiple microphones results

- in a maximum search on the sum or mean of the squared microphone cepstra for $\widehat{\mathrm{E}}\{\phi_{rq_0}\} = \phi_{rq_0}$

$$q_0^{\mathrm{ML},R=0} = \arg\max_q \sum_{m=1}^{M} \phi_{q,m}^2\,, \tag{2.50}$$

- in a maximum search on the sum or mean of the microphone cepstra for $\widehat{\mathrm{E}}\{\phi_{rq_0}\} = |\phi_{rq_0}|$

$$q_0^{\mathrm{ML},R=0} = \arg\max_q \sum_{m=1}^{M} \phi_{q,m}\,, \tag{2.51}$$

- approximately in a maximum search on the quefrency smoothed cepstrum for $\widehat{\mathrm{E}}\{\phi_{rq_0}\} = |\check{\phi}_{rq_0}|$

$$\begin{aligned} q_0^{\mathrm{ML}} &= \arg\max_q \sum_{m=1}^{M} |\check{\phi}_{q,m}| \left(2\phi_{q,m} - |\check{\phi}_{q,m}|\right) \\ &\approx \arg\max_q \sum_{m=1}^{M} \check{\phi}_{q,m}\,. \end{aligned} \tag{2.52}$$

We refer to these approaches as Multi-Microphone Cepstral ML fundamental period estimator (MM-CML). Another approach that exploits the information of multiple microphones is to apply a ML fundamental period estimation on the output of a beamformer (Beamforming based Cepstral ML fundamental period estimator (BF-CML)). The output of a beamformer has an increased signal-to-noise ratio as compared to each single microphone channel. This results in more prominent spectral harmonics and thus in an increased cepstral peak. Under the Gaussian assumption, the variance of the non-speech cepstral coefficients stays unchanged, as it is independent of the signal power.

While the coefficients rq_0 of the microphone cepstra are correlated, the remaining coefficients can be assumed to be uncorrelated between microphones. Thus, adding the microphone cepstra as proposed by the MM-CML estimators of Section 2.4.3 increases the difference between the cepstral peak and the cepstral variance more directly. While both approaches, MM-CML and BF-CML, increase the estimation performance, the superiority of the cepstral averaging approach is demonstrated in Section 2.4.5.

2.4.4 MAP fundamental period tracking

To decrease the amount of estimation errors, the fundamental period can be tracked over time. To achieve this, we extend the ML fundamental period estimator towards a MAP fundamental period estimator similar to [Droppo and Acero, 1998, Tabrikian *et al*, 2004]. The MAP fundamental period estimator is given by

$$q_0^{\mathrm{MAP}} = \arg\max_{q_0} \left(p(q_0) p(\mathbf{\Phi}|q_0) \right) . \tag{2.53}$$

Thus, in addition to the likelihood (2.49), (2.39) we also need to model the *a priori* probability of the fundamental period $p(q_0)$. As proposed in [Tabrikian *et al*, 2004], we incorporate the information of Λ consecutive frames by treating the *a priori* fundamental period probability $p(q_0)$ as a first order Markov chain, as

$$p(q_0(l)) = \prod_{\lambda=0}^{\Lambda-1} p(q_0(l-\lambda)|q_0(l-\lambda-1)) , \tag{2.54}$$

where $q_0(l)$ are the states and $p(q_0(l)|q_0(l-1))$ is the transition probability density function. For the initial state we choose $p(q_0(l-\Lambda+1)|q_0(l-\Lambda)) = 1$. The transition probability density function can be chosen to be Gaussian, *i.e.*

$$p(q_0(l)|q_0(l-1)) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathrm{tracking}}}} \exp\left(-\frac{\left(q_0(l) - q_0(l-1)\right)^2}{2\sigma^2_{\mathrm{tracking}}} \right) ,$$

whereas the standard deviation $\sigma_{\mathrm{tracking}}$ can be found using labelled training data, *e.g.* the data of the Keele database [Meyer, Accessed 2006].

Similar to (2.49) we assume that the cepstral coefficients of consecutive signal segments, *given* q_0, are independent. Then, the MAP estimator including the information of the last Λ signal segments is given by [Tabrikian *et al*, 2004]

$$\begin{aligned} \mathbf{q_0}^{\mathrm{MAP}}(l) &= \arg\max_{\mathbf{q_0}} \prod_{\lambda=0}^{\Lambda-1} p(\mathbf{\Phi}(l-\lambda)|q_0(l)) \; p(q_0(l)|q_0(l-\lambda-1)) \\ &= \arg\max_{\mathbf{q_0}} \sum_{\lambda=0}^{\Lambda-1} \log\Big(p(\mathbf{\Phi}(l-\lambda)|q_0(l))\Big) + \log\Big(p(q_0(l)|q_0(l-\lambda-1))\Big) , \end{aligned} \tag{2.55}$$

where $\mathbf{q_0}^{\mathrm{MAP}}(l) = \{q_0(l), ..., q_0(l-\Lambda+1)\}$ is the sequence of fundamental period estimates that is optimal in the MAP sense.

While (2.55) requires estimating the whole sequence $\mathbf{q_0}^{\mathrm{MAP}}(l)$ at each segment l, the estimator is simplified if the estimate at segment l is based on the MAP estimates $q_0^{\mathrm{MAP}}(l-\lambda-1)$ of previous frames similar to [Tabrikian *et al*, 2004], as

$$q_0^{\mathrm{MAP}}(l) = \arg\max_{q_0} \sum_{\lambda=0}^{\Lambda-1} L_{q_0}(\mathbf{\Phi}(l-\lambda)) + B_{q_0}\Big(q_0^{\mathrm{MAP}}(l-\lambda-1)\Big)\,, \tag{2.56}$$

with the log likelihood

$$L_{q_0}\left(\mathbf{\Phi}(l-\lambda)\right) = \log\Big(p(\mathbf{\Phi}(l-\lambda)|q_0)\Big)\,, \tag{2.57}$$

and the logarithmic transition probability density function

$$B_{q_0}\Big(q_0^{\mathrm{MAP}}(l-\lambda-1)\Big) = \log\Big(p(q_0|q_0^{\mathrm{MAP}}(l-\lambda-1))\Big)\,. \tag{2.58}$$

As the pitch tracking algorithm is meant to provide pitch estimates for low-delay applications, no major look ahead is possible and an instantaneous decision is needed in each signal segment. A drawback of (2.56) is that all segments $\{l, ..., l-\Lambda+1\}$ contribute equally to the current estimate at segment l. To emphasize the information in recent signal segments, instead of using (2.56), we propose to realize (2.55) via a recursive averaging as

$$W_{q_0}(l) = \alpha W_{q_0}(l-1) + (1-\alpha)\Big(B_{q_0}\Big(q_0^{\mathrm{MAP}}(l-1)\Big) + L_{q_0}(\mathbf{\Phi}(l))\Big) \tag{2.59}$$

using the initializion $W_{q_0}(0) = L_{q_0}(\mathbf{\Phi}(0))$ and the MAP fundamental period estimate

$$q_0^{\mathrm{MAP}}(l) = \arg\max_{q_0} W_{q_0}(l)\,. \tag{2.60}$$

2.4.5 Experimental results

We compare the MM-CML estimators based on the summation of the microphone cepstra proposed in Section 2.4.3 to a cepstral ML estimation on the output signal of a beamformer (BF-CML). Further, we give the results for the Multi-Microphone Cepstral MAP fundamental period estimator (MM-CMAP) for $R = 0$ and $R = 1$. For the evaluation we use the Keele database [Meyer, Accessed 2006] that consists of 5 male and 5 female speakers and up to 40 s of speech per speaker. The sampling rate is $f_{\mathrm{s}} = 20\,\mathrm{kHz}$, the segment size 25.6 ms and the frame shift 10 ms. This corresponds to $N = 512$ and $L = 200$ in (1.1). We choose a rectangular spectral analysis window w_n in (1.1) and assume that the cepstral variance is quefrency independent. Further, we choose $q_{\mathrm{low}} = 40$ (2 ms) in (2.39). For the MAP algorithm we choose the smoothing constant $\alpha = 0.8$ in (2.59). The standard deviation of the *a priori* probability is determined based on

the labelled training data [Meyer, Accessed 2006] and set to $\sigma_{\text{tracking}} = 23$ bins which corresponds to 1.1 ms.

To decouple the evaluation of the fundamental period estimators from the problem of automatic voiced/unvoiced classification, a fundamental period estimation is applied only on those signal segments that are marked as voiced in the Keele database. The estimated fundamental frequency $\hat{f}_0$ is compared to the reference fundamental frequency f_0 of the Keele database. For the evaluation we use the Gross Error Rate (GER) and the relative Root Mean Square Error (RMSE) according to [Flego, 2006]. The GER is given as the percentage of signal segments that have a fundamental frequency estimate that deviates by more than $\theta\%$ of the reference fundamental period.

$$\text{GER}(\theta) = \frac{1}{N_\text{v}} \sum_{l=1}^{N_\text{v}} \left\{ \frac{|\hat{f}_0(l) - f_0(l)|}{f_0(l)} > \theta\% \right\}, \tag{2.61}$$

where N_v is the number of voiced signal segments. The relative RMSE

$$\text{RMSE}(\theta) = \sqrt{\frac{1}{N_\theta} \sum_{l \in \Omega(\theta)} \left(\frac{\hat{f}_0(l) - f_0(l)}{f_0(l)} \right)^2}. \tag{2.62}$$

is evaluated only for those N_θ signal segments of the set $\Omega(\theta)$, which have a relative fundamental frequency estimation error smaller than $\theta\%$. It can be seen as a measure for the fine fundamental frequency estimation error [Flego, 2006].

For the evaluation, we generate ten microphone signals with stationary diffuse additive white Gaussian noise at several segmental SNRs. The diffuse noise field is created as detailed *e.g.* in [Habets *et al*, 2008]. The ten microphones are assumed to be linearly spaced with a 5 cm gap between each microphone. For the BF-CML approach, the ten microphone signals are summed in time domain and the ML estimator is applied on the cepstrum of the sum. We thus simulate the case of a source at the broadside of the array with its location perfectly known. For the MM-CML algorithm, the cepstrum is computed for each microphone signal, and the maximum of the sum of the cepstra is searched as proposed in Section 2.4.3. Note that for the MM-CML a source localization is not needed if the maximal time delay between the microphone signals (Maximum Microphone Distance)/(340 m/s) is small as compared to the segment length N/f_s, as the phase of the complex spectra is neglected when computing the cepstrum via (2.1).

In Figure 2.12 we present the results when the peak of the squared cepstrum is searched for as given in (2.41), in Figure 2.13 we search for a positive peak according to (2.44), and in Figure 2.14 we smooth the cepstrum and use (2.47). For the smoothing kernel $w_{\text{H},q}$, we use a normalized Hamming window of length $f_\text{s}/(2000\,\text{Hz}) = 10$. It can be

clearly seen that searching for a positive peak increases the robustness of the fundamental period estimator in terms of the GER. The quefrency smoothing of the cepstrum (Figure 2.14) increases the performance even more. In all cases of figures 2.12, 2.13, and 2.14 the proposed MM-CML approaches of Section 2.4.3 clearly outperforms a delay-and-sum beamformer approach BF-CML in terms of the GER and the RMSE. As now the cepstral transform has to be computed M times, the increased performance goes along with an increased computational complexity. When the fundamental period is tracked over time (MM-CMAP), the results are further enhanced in terms of a lower GER and the estimation performance can be seen to be much more robust in noisy environments. For Figure 2.13 and Figure 2.14 also the fine fundamental frequency estimation error (RMSE) decreases when the MAP fundamental period tracking is employed. In case a maximum is searched on the squared cepstrum, the performance in terms of the RMSE decreases when the fundamental period tracking is used, while the GER indicates an increased performance (cf. Figure 2.12). For the MM-CMAP we also present the results for $R = 1$. It can be seen that incorporating a rahmonic reduces the fine pitch estimation error at the price of more outliers in terms of the GER. As the rahmonics are often much smaller than the fundamental period peak [Noll, 1967], it may happen that the sum of two noise bins is larger than the sum of the fundamental period peak and its rahmonic. Additionally, especially for male speakers, incorporating the rahmonics increases fundamental period halving errors. In that cases, estimation errors occur that result in an increased GER.

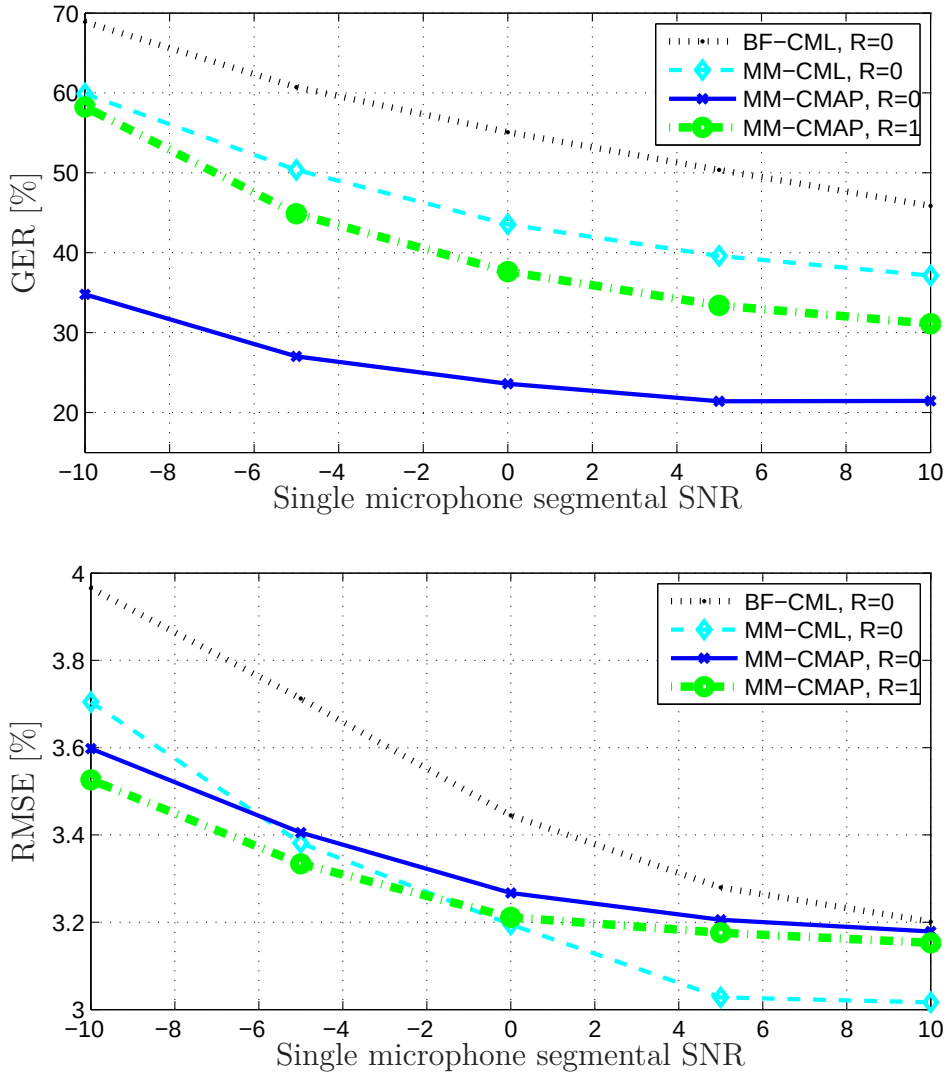


Figure 2.12: GER (upper panel) and RMSE (lower panel) for various input segmental SNRs for $\theta = 10\%$, $M = 10$, and diffuse white Gaussian noise. The proposed MM-CML approach outperforms a delay-and-sum-beamformer (BF-CML) in terms of GER and RMSE for all considered input SNRs. Maximum a posteriori fundamental period tracking (MM-CMAP) further enhances the estimation performance. Here, a peak on the squared cepstrum is searched as proposed in (2.41).

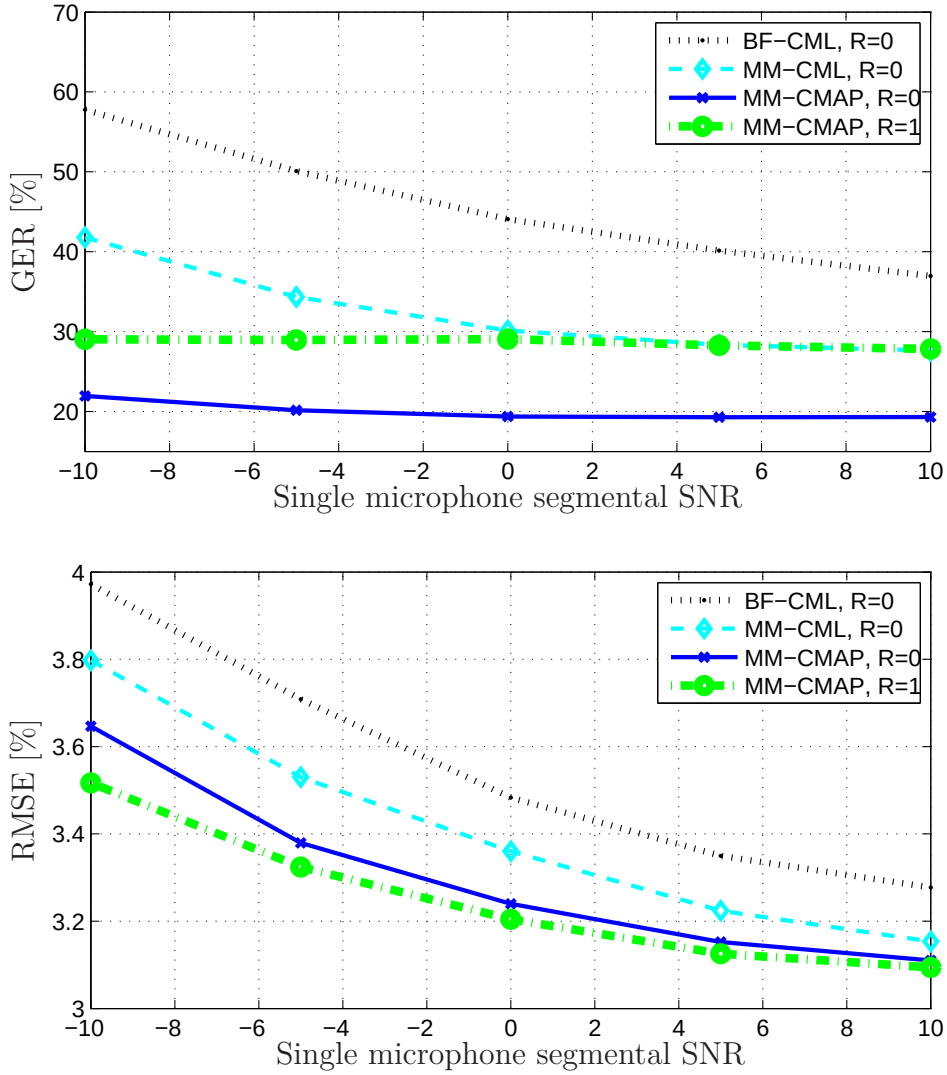


Figure 2.13: As Figure 2.12, but a positive peak is searched, as proposed in (2.44). As compared to Figure 2.12, where the maximum of the squared cepstrum is searched, the results indicate an increased robustness.

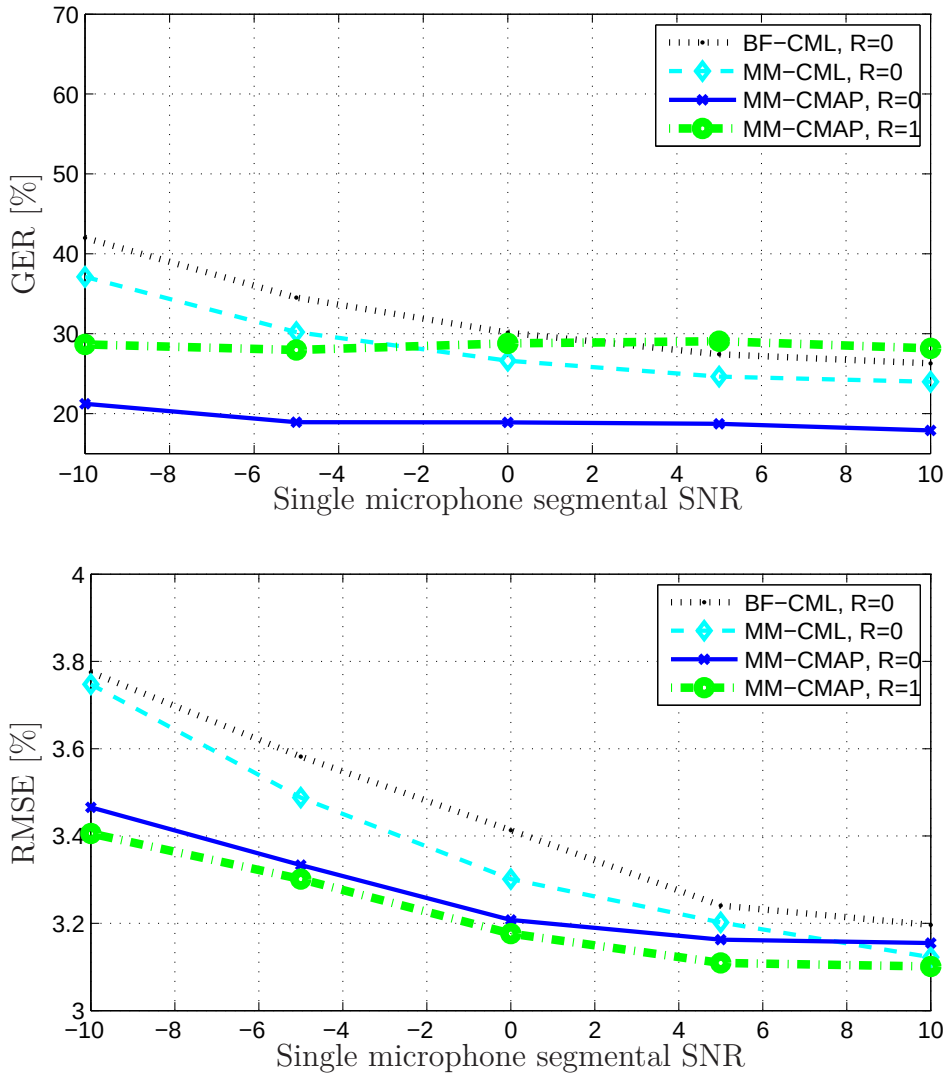


Figure 2.14: As Figure 2.13, but the cepstrum is smoothed over quefrency resulting in (2.47). The additional smoothing increases the robustness further (cf. figures 2.13 and 2.14).

2.5 Conclusions

In this chapter the properties and applications of the cepstrum are discussed. Speech is shown to be very compactly represented in the cepstral domain. For speech processing algorithms this compactness is a desirable property as it allows for a selective treatment of speech related coefficients and the remaining coefficients. The concepts of Temporal Cepstrum Smoothing (TCS) and Cepstral Nulling (CN) are introduced where it is proposed to apply little smoothing or no modification to the speech related cepstral coefficients, while strongly smoothing or nulling the remaining coefficients. However, due to the logarithmic compression inherent in the cepstral transform, a modification in the cepstral domain results in a signal power bias in the spectral domain.

An explicit expression is derived to account for the signal power bias that occurs when a spectral quantity is modified in the cepstral domain. The bias compensation is based on the analysis of the statistical properties of cepstral coefficients. If χ-distributed spectral amplitudes are smoothed in the cepstral domain, the resulting smoothed spectral amplitudes are found to be also approximately χ-distributed but with more degrees of freedom and less signal power. Explicit expressions for the mean, the variance, and the covariance of cepstral coefficients and the logarithmic periodogram are derived, parameterized by the shape parameter of χ-distributed spectral amplitudes. The spectral correlation introduced by tapered spectral analysis windows is shown to result in a decline of the cepstral variance for an increasing cepstral index. The key finding for the proposed bias compensation is that the degrees of freedom of χ-distributed spectral amplitudes are directly related to their average cepstral variance. Thus, for a given modification of cepstral coefficients, the shape parameter of the spectral amplitudes after the modification is determined. Finally, an expression for the bias compensation is derived that is only dependent on the shape parameters before and after the cepstral modification. As the parameterized χ-distribution for the spectral amplitudes is assumed, the presented results hold for Gaussian, super-Gaussian, and slightly sub-Gaussian distributed complex spectral coefficients. The proposed bias compensation method is computationally inexpensive and shown to work very well for white and colored signals, as well as for rectangular and tapered spectral analysis windows.

To determine the set of cepstral coefficients that represent speech, the determination of the speech fundamental period is necessary. Maximum likelihood and maximum a posteriori estimators for a fundamental period estimation in the cepstral domain are derived, which also motivate the well known approach by Noll [Noll, 1967]. For spectrally uncorrelated data a maximum search is found to be optimal in the maximum likelihood sense. When extending the likelihood function towards multiple microphones, the maximum likelihood solution results in a maximum search on the sum of all microphone cepstra. This approach is shown to outperform a cepstral maximum search on the cepstrum of the output of a delay-and-sum beamformer in terms of the Gross Error Rate (GER) and

the Root Mean Square Error (RMSE) for all considered signal-to-noise ratios at the cost of an increased computational complexity. Finally, the maximum likelihood estimator is extended to a Maximum *A Posteriori* (MAP) fundamental period tracking that substantially improves the robustness in noisy environments.

Chapter 3

Temporal Cepstrum Smoothing for Speech Enhancement

In this chapter we address the Temporal Cepstrum Smoothing (TCS) approach, introduced in Section 2.2 and summarized in Algorithm 1.

In TCS a spectral quantity is transformed into the cepstral domain, selectively smoothed over time and transformed back into the spectral domain. The selective smoothing is done via the quefrency dependent smoothing constant α_q, which is close to zero for the speech related cepstral coefficients $q \in \mathbb{Q}$ and close to one for the remaining coefficients $q \in \overline{\mathbb{Q}}$. For the determination of the speech related cepstral coefficients $\mathbb{Q} = \{0, \ldots, q_{\text{low}}, \mathbb{Q}_{\text{pitch}}\}$, defined in (2.2), we need to determine the cepstral coefficients $\mathbb{Q}_{\text{pitch}}$ that represent the fundamental period.

Since the power of voiced sounds is less at high frequencies, *e.g.* [Loizou, 2007, Section 4.2], the estimation of the fundamental period is more robust if only the spectrum up to a certain cut-off frequency is considered. This low-pass filtering of the log-spectrum can be achieved by convolving each cepstral frame with a short Hamming window, $w_{\text{H},q}$, of length τ_{H} taps as proposed in (2.45)

$$\check{\phi}_q = \phi_q * w_{\text{H},q}\,. \tag{3.1}$$

We found that a simple maximum search on the smoothed cepstrum, as derived in (2.48), leads to sufficiently robust fundamental period estimates for cepstral smoothing. The cepstral index $q_0(l)$ that most likely represents f_0 is thus found as

$$q_0(l) = \underset{q}{\text{argmax}}\Big\{\check{\phi}_q(l) | q_{\text{low}} \leq q \leq q_{\text{high}}\Big\}\,, \tag{3.2}$$

where the search is limited to possible fundamental frequencies between $f_{0,\text{low}}$ and $f_{0,\text{high}}$, resulting in the range $q_{\text{low}} = \lfloor f_{\text{s}}/f_{0,\text{high}} \rfloor$ to $q_{\text{high}} = \lfloor f_{\text{s}}/f_{0,\text{low}} \rfloor$, with f_{s} the sampling rate and $\lfloor \cdot \rfloor$ the flooring operator towards the nearest integer number.

Note that (3.2) only yields meaningful results if voiced speech is present. To detect voiced speech sounds, we compare the found peak value to a threshold, Λ^{thr}. Thus, the set of cepstral bin indices associated with the fundamental frequency, $\mathbb{Q}_{\text{pitch}}$, is gained as

$$\mathbb{Q}_{\text{pitch}} = \begin{cases} \{q_0 - \Delta q_0, ..., q_0 + \Delta q_0\} & \text{if } \check{\phi}_q(l) \geq \Lambda^{\text{thr}} \\ \emptyset & \text{else}\,, \end{cases} \tag{3.3}$$

where $q \in \{q_0 - \Delta q_0, ..., q_0 + \Delta q_0\}$ is the range of cepstral bins that most likely represent the fundamental period, Δq_0 is a small margin, and $\emptyset$ is the empty set. A suitable value for the threshold Λ^{thr} is found from tests with representative noisy data. Note that for cepstral smoothing Λ^{thr} is not a sensitive parameter, as long as it is chosen rather low.

To preserve the speech spectral envelope, less smoothing is applied to the cepstral coefficients $q \leq q_{\text{low}}$ than to the coefficients $q \in \overline{\mathbb{Q}}$. To achieve this, a smoothing constant α_q^{const} is chosen to gradually increase with increasing q. The exact choice for α_q^{const} is given in the respective section of this chapter. To avoid a strong smoothing of the speech fundamental period peak, an adaptive smoothing factor $\alpha_q(l)$ is gained as

$$\alpha_q(l) = \begin{cases} \alpha_{\text{pitch}} & \text{, if } q \in \mathbb{Q}_{\text{pitch}}\,, \\ \beta\,\alpha_q(l-1) + (1-\beta)\,\alpha_q^{\text{const}} & \text{, else}\,, \end{cases} \tag{3.4}$$

where α_{pitch} is a value rather close to zero. The smoothing constant β is a forgetting factor that determines how fast the value of $\alpha_q(l)$ rises back from α_{pitch} to α_q^{const}, if it has been lowered in previous frames. Due to (3.4), a detection error of the fundamental period in the current frame l does not lead to an immediate strong smoothing of the cepstral fundamental period bin in step 4 of Algorithm 1.

3.1 Smoothing spectral gain functions

In speech enhancement, an estimate of clean speech is often obtained by multiplying the noisy speech with a spectral gain function in the Discrete Fourier Transform (DFT) domain as given in (1.3). However, especially for a high frequency resolution that enables the suppression of noise between spectral harmonics of voiced sounds, spectral outliers in the spectral gain function lead to musical noise.

One way of preventing spectral outliers in single channel speech enhancement gain functions is to combine the gain function for clean speech estimation $G_{\mathcal{H}_{1,k},k}$ with an *a posteriori* Speech Presence Probability (SPP) estimate $P(\mathcal{H}_{1,k}|\, Y_k, \sigma_{\text{S}}^2, \sigma_{\text{N}}^2)$ as given in

(1.3) [Malah *et al*, 1999]. Alternatively in [Linhard and Haulick, 1999] a filter is described that has several parameters for adapting the spectral gain function to the noise condition. A third strategy is to search and remove spectral peaks in the filtered signal that lead to musical noise [Goh *et al*, 1998]. In [Gustafsson *et al*, 2001] a recursive averaging is applied to the spectral gain function that smoothes out fluctuations, while in [Esch and Vary, 2009, Brandt and Bitzer, 2009] a smoothing along frequency is proposed. As such a smoothing in the time-frequency domain may also severely affect speech components, the smoothing constant of these algorithms have to be carefully adapted.

In this section we propose to temporally smooth the spectral gain function in the cepstral domain to reduce spectral outliers that may yield musical noise. This is a very flexible application for TCS as it can be used for any speech enhancement algorithm that obtains a clean speech spectral estimate by applying a multiplicative spectral gain function or a binary mask as $\widehat{S}_k = G_k Y_k$. This approach has been introduced for the case of single channel speech enhancement [Breithaupt *et al*, 2007] and has then been carried over to blind source separation [Madhu *et al*, 2008].

In single channel speech enhancement, the gain function may be given by the Wiener Filter (1.4). As the Wiener filter is a function of the speech and noise spectral power, and thus of squared spectral quantities, we define the cepstrum as given in (2.1) where Φ_k represents the spectral gain function $\widetilde{G}_k$ which is limited to be larger than $G_{\min}$ according to (1.7). The spectral gain function is then smoothed via Algorithm 1.

3.1.1 Bias compensation

For the computation of the bias correction $\mathcal{B}$ in Algorithm 1, we may not use Algorithm 4, as the bias correction in Algorithm 4 relies on the assumption that the spectral quantity Φ_k is χ^2-distributed. In this section, we aim at deriving a bias correction that holds for arbitrary spectral gain functions, *e.g.* the Wiener filter, the Log Spectral Amplitude (LSA) estimator [Ephraim and Malah, 1985] or spectral masks for blind source separation [Madhu *et al*, 2008]. Therefore, a very general assumption is made on the distribution of the spectral gain function: we assume that the gain function G_k is uniformly distributed between 0 and 1, and then limited to be larger than $G_{\min}$ as given in (1.7). The results of this section are partly presented in [Gerkmann *et al*, 2008a].

With Heaviside's step function

$$\Theta(b) = \begin{cases} 0 & \text{if } b < 0 \\ 1 & \text{if } b \geq 0 \end{cases},$$

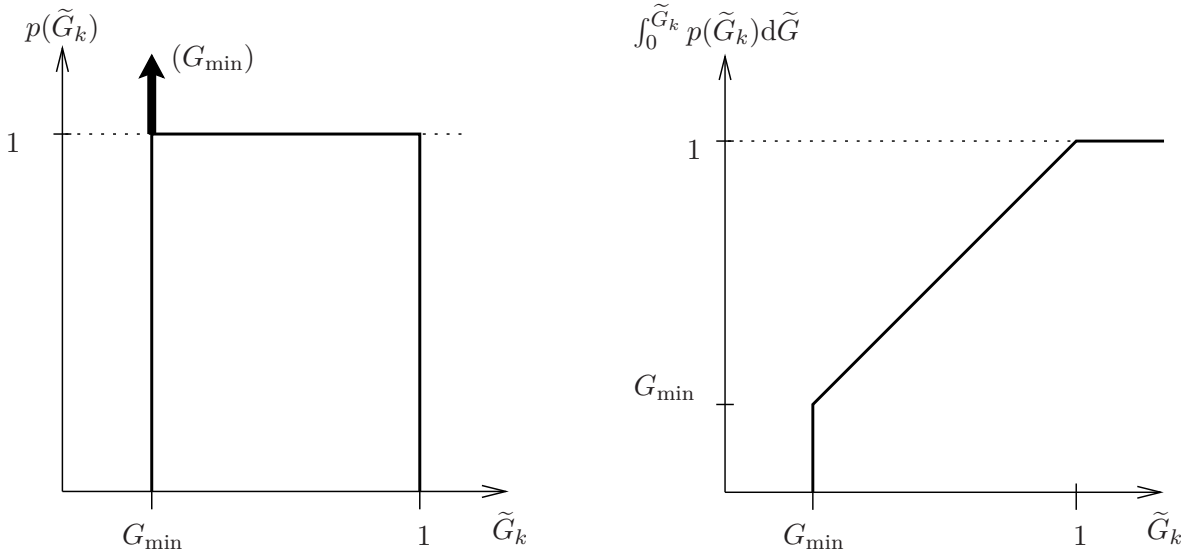


Figure 3.1: The assumed distribution $p(\widetilde{G}_k)$ of the gain function (left) and its cumulative distribution (right).

Dirac's delta function $\delta(\cdot)$ is given by

$$\int_{-\infty}^{b} \delta(a)\mathrm{d}a = \Theta(b)\,.$$

Using Heaviside's step function and Dirac's delta function, the probability density function of the limited gain function $\widetilde{G}_k$ can be written as

$$p(\widetilde{G}_k) = \Theta(\widetilde{G}_k - G_{\min}) - \Theta(\widetilde{G}_k - 1) + G_{\min}\,\delta(\widetilde{G}_k - G_{\min})\,, \tag{3.5}$$

such that

$$\int_0^{\widetilde{G}_k} p(\widetilde{G}_k)\mathrm{d}\widetilde{G}_k = \begin{cases} 1 & \text{, if } \widetilde{G}_k > 1 \\ \widetilde{G}_k & \text{, if } G_{\min} \leq \widetilde{G}_k \leq 1 \\ 0 & \text{, otherwise.} \end{cases} \tag{3.6}$$

The resulting distribution is visualized in Figure 3.1.

To derive the bias $\mathcal{B}$ in Algorithm 1, we now assume that the cepstral smoothing perfectly approximates the expected value operator. However, due to the logarithm in the definition of the cepstrum (2.1) this expectation is not taken in the linear domain, but in the log domain, and thus results in the bias

$$\mathcal{B} = \frac{\mathrm{E}\big\{\widetilde{G}_k\big\}}{\exp\Big(\mathrm{E}\big\{\log\big(\widetilde{G}_k\big)\big\}\Big)}\,. \tag{3.7}$$

With the distribution given in (3.5) the expectations in (3.7) result in

$$\begin{aligned}\mathrm{E}\Big\{\widetilde{G}_k\Big\} &= \int_{G_{\min}}^{1} \widetilde{G}_k \mathrm{d}\widetilde{G}_k + \int_{-\infty}^{\infty} \widetilde{G}_k G_{\min}\, \delta(\widetilde{G}_k - G_{\min}) \mathrm{d}\widetilde{G}_k \\ &= \frac{1}{2}\left(1 - G_{\min}^2\right) + G_{\min}^2 \\ &= \frac{1}{2}\left(1 + G_{\min}^2\right) , \end{aligned} \tag{3.8}$$

and with [Gradshteyn and Ryzhik, 2000, (2.723.1)] we find

$$\begin{aligned}\mathrm{E}\Big\{\log \widetilde{G}_k\Big\} &= \int_{G_{\min}}^{1} \log\Big(\widetilde{G}_k\Big)\, \mathrm{d}\widetilde{G}_k + \int_{-\infty}^{\infty} \log\Big(\widetilde{G}_k\Big)\, G_{\min}\, \delta(\widetilde{G}_k - G_{\min}) \mathrm{d}\widetilde{G}_k \\ &= \Big[\widetilde{G}_k \log\Big(\widetilde{G}_k\Big) - \widetilde{G}_k\Big]_{G_{\min}}^{1} + G_{\min} \log(G_{\min}) \\ &= G_{\min} - 1 \,. \end{aligned} \tag{3.9}$$

With (3.8) and (3.9), the ratio (3.7) results in:

$$\mathcal{B} = \frac{1}{2}\left(1 + G_{\min}^2\right) \mathrm{e}^{1 - G_{\min}} \,. \tag{3.10}$$

The bias correction factor $\mathcal{B}$ is then applied to the smoothed spectral gain function in step 6 of Algorithm 1.

To check the plausibility of (3.10), note that the numerator of (3.7) may be seen as an arithmetic mean of the gain function which is always less or equal to the denominator which corresponds to the geometric mean. Then, the bias $\mathcal{B}$ can be seen as the ratio between arithmetic and geometric mean, which increases with a decreasing $G_{\min}$. This can be observed in Figure 3.2, where the bias correction $\mathcal{B}$ is plotted as a function of $G_{\min}$. For $G_{\min} = 1$, the gain function is a constant which is the only case where geometric mean and arithmetic mean are equivalent and the bias correction factor is 1.

3.1.2 Experimental results

We now evaluate the proposed TCS of spectral gain functions. For the evaluation, the probability of speech presence in (1.3) is assumed to be $P(\mathcal{H}_{1,k} | \, Y_k, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{N}}^2) = 1$ for all time-frequency points. The estimation of the probability of speech presence $P(\mathcal{H}_{1,k} | \, Y_k, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{N}}^2)$ is treated separately in Chapter 5. The gain function G_k is given by the Wiener filter, *i.e.* $G_k = G_{\mathcal{H}_{1,k},k} = \frac{\xi_k}{1+\xi_k}$. The gain function is first limited to be larger than $20\log_{10}(G_{\min,1}) = -22\,\mathrm{dB}$ and after TCS limited to be larger than $20\log_{10}(G_{\min}) = -17\,\mathrm{dB}$. The bias correction is computed with the first limit $20\log_{10}(G_{\min,1}) = -22\,\mathrm{dB}$. While the spectral noise power is estimated using the minimum statistics approach

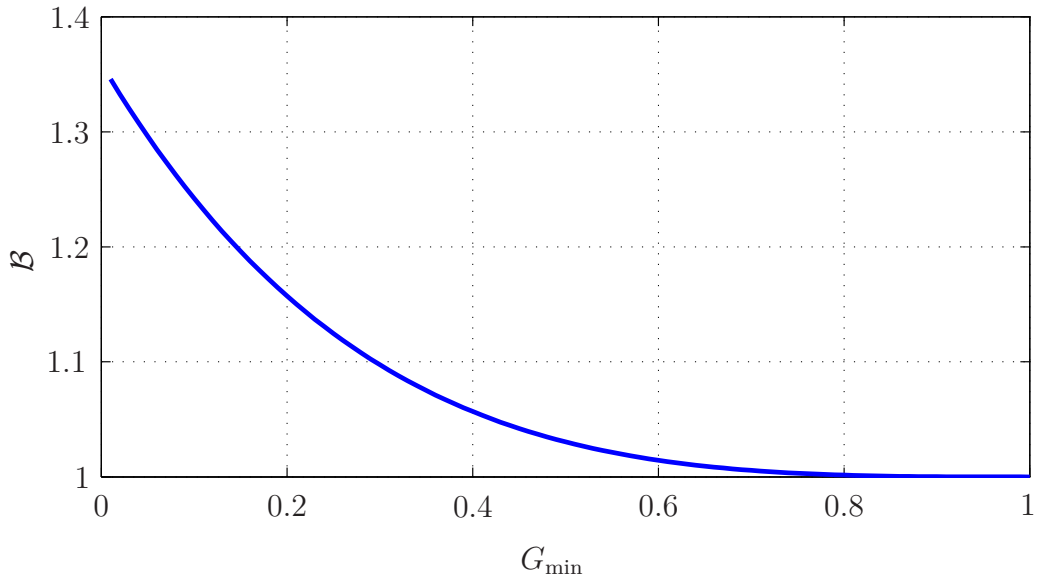


Figure 3.2: The bias correction $\mathcal{B}$ for a TCS of the filter gain G_k, as a function of the lower limit $G_{\min}$ of the gain function.

[Martin, 2001], the *a priori* Signal-to-Noise Ratio (SNR) is estimated using the decision-directed approach (3.14) with $\alpha_{\mathrm{dd}} = 0.94$. For the short-time Fourier analysis (1.1) we use Hann windows w_n with a length of 32 ms and 50% overlap.

For the recursive smoothing constant in (3.4), we choose:

$$\alpha_q^{\mathrm{const}} = \begin{cases} 0 & , q \in \{0, ..., 2\} \\ 0.3 & , q \in \{3, ..., 6\} \\ 0.5 & , q \in \{7, ..., 12\} \\ 0.6 & , q \in \{13, ..., 19\} \\ 0.97 & , q \in \{20, ..., 256\}\,, \end{cases} \tag{3.11}$$

The remaining parameters used for TCS of spectral gain functions are summarized in Table 3.1.

We compare the cepstral smoothing of spectral gain functions to an approach where we choose a large value $\alpha_{\mathrm{dd}} = 0.98$ for the decision-directed SNR estimation approach (3.14), as increasing α_{dd} is known to reduce the musical noise phenomenon when used with estimators for the clean speech spectral amplitudes [Cappé, 1994]. With these settings, neither the approach with $\alpha_{\mathrm{dd}} = 0.98$ nor the approach with $\alpha_{\mathrm{dd}} = 0.94$ and the TCS of the spectral gain function yield musical noise for white stationary noise.

Smoothing factor for $\mathbb{Q}_{\text{pitch}}$ (3.4)	α_{pitch}	$= 0.2$
Threshold for voiced/unvoiced decision (3.3)	Λ^{thr}	$= 0$
Lower bound for the q_0 search (3.2)	q_{low}	$= \left\lfloor \frac{f_{\text{s}}}{300\,\text{Hz}} \right\rfloor$
Upper bound for the q_0 search (3.2)	q_{high}	$= \left\lfloor \frac{f_{\text{s}}}{70\,\text{Hz}} \right\rfloor$
Margin for $\mathbb{Q}_{\text{pitch}}$ (3.3)	Δq_0	$= 2$
Length of the cepstral low-pass (3.1)	τ_{H}	$= f_{\text{s}}/2000\,\text{Hz}$
Smoothing constant for (3.4)	β	$= 0.96$
Lower bound on the gain function before TCS	$20\log_{10}(G_{\text{min},1})$	$= -22\,\text{dB}$
Lower bound on the gain function after TCS	$20\log_{10}(G_{\text{min}})$	$= -17\,\text{dB}$
Sampling rate	f_{s}	$= 16\,\text{kHz}$

Table 3.1: Parameters for TCS of the spectral gain function.

We process 320 samples from the TIMIT database [Garofolo, 1988, dialect region 6] that sum up to approximately 15 minutes of fluent, phonetically balanced conversational speech of both male and female speakers. The speech samples are disturbed by stationary white Gaussian noise, nonstationary traffic noise at a crowded street, and babble noise in a restaurant for input segmental SNRs between -5 and 15 dB. The improvement of the segmental SNR, the segmental speech SNR, and the segmental noise reduction [Breithaupt, 2008, Lotter, 2004, Lotter and Vary, 2005] are given in Figure 3.3. For all three measures higher values indicate an increased performance. In particular, a higher speech SNR indicates less speech distortions. The segmental SNR considers both speech distortions and noise reduction. For input SNRs below 0 dB the segmental SNR would indicate an improvement even if the gain function is zero for all time-frequency points. Therefore it has to be read together with the segmental speech SNR. It can be seen that in terms of the instrumental measures, the approaches yield rather similar results. The approach with TCS yields slightly less noise reduction, a higher speech SNR and for white noise a slightly higher SNR improvement as compared to the approach with $\alpha_{\text{dd}} = 0.98$.

Per definition, the cepstrum of a spectral quantity is given by the inverse Fourier transform of the *logarithm* of the spectral quantity (2.1). If the spectral quantity is the speech spectral power, the logarithm is important for the compression of the spectral harmonics. If no logarithm is used, the lower spectral harmonics would usually be much larger than the higher harmonics and hence would not be mapped to a strong peak in the cepstrum. However, in contrast to the speech spectral power, the gain function resulting from a Wiener filter or a binary mask for blind source separation is usually bound to

values between $G_{\min}$ and 1. Therefore, an additional compression by the log function is not mandatory for the selective cepstral smoothing of spectral gain functions. One advantage of using no logarithmic compression is that the selective smoothing process results in unbiased smoothed spectral quantities. In Figure 3.3 we also present the results when no logarithmic compression is applied. It can be seen that this results in a lower speech SNR. While an increased noise reduction may be observed when no logarithmic compression is used, especially at high input SNRs the segmental SNR is larger when the logarithmic compression is applied. Further, informal listening shows that using the logarithmic compression reduces noise shaping effects caused by the recursive smoothing. Noise shaping effects occur after a speech sound has ended as then, due to the recursive smoothing, the gain function may still be large although it should be zero. This effect is reduced if the logarithmic compression is applied, as then small values have a strong effect on the averaging and, compared to the case when no logarithmic compression is used, the smoothed gain function will decrease quicker after a speech sound has ended.

While the instrumental measures indicate a similar performance in terms of speech distortions, noise reduction and SNR improvement, the major benefit of TCS can be seen in the reduction of processing artifacts such as spectral outliers that may be perceived as musical noise. While processing artifacts may considerably reduce the perceived signal quality, none of the used measures is designed to predict artifacts such as musical noise. In Figure 3.4 the spectrogram of the exemplary sentence "Please shorten this skirt for Joice" is given, which is spoken by a female. In Figure 3.5 the spectrograms for noisy speech disturbed by stationary white Gaussian noise at 0 dB segmental SNR are given, while in Figure 3.6 spectrograms for nonstationary babble noise are given. With respect to the noisy signal, after speech enhancement the background noise is reduced. As we limit the gain function to be larger than $G_{\min} = -17\,\text{dB}$ to reduce processing artifacts and speech distortions as discussed in Chapter 1, no complete cancellation of the noise is achieved. Comparing TCS in figures 3.5(b) and 3.6(b) to the approach without TCS in figures 3.5(c) and 3.6(c), it can be seen that TCS notably preserves plosives (*e.g.* at 2.1 s), vowels (*e.g.* at 2.3 s) and fricatives (*e.g.* at 4.2 s for white noise). Most importantly, it may be observed in Figure 3.6 that for babble noise the residual noise signal is much smoother for the approach with TCS, *i.e.* spectral outliers are strongly reduced. This impression is also confirmed by informal listening, where a strong reduction of the musical noise phenomenon may be observed.

The most reliable way to assess the signal quality is to conduct listening experiments [Dreiseitel and Schmidt, 2006]. Therefore, in [Breithaupt *et al*, 2007] we conducted listening experiments for a cepstral smoothing of spectral gain functions. Even though the parameter setting in the listening experiments were slightly different from those recommended in Table 3.1, we believe that the listening experiments prove the general strengths of TCS based approaches, namely a considerable reduction of spectral outliers that result in more natural sounding residual noise and a higher signal quality. In

the listening experiments we compare the performance of the decision-directed SNR estimation approach and an *a posteriori* SPP estimator according to [Malah *et al*, 1999] to a TCS of spectral gain functions. We use nonstationary babble noise, nonstationary subway noise and stationary pink Gaussian noise. For each noise type, ten different speech samples from [Garofolo, 1988] were presented, five spoken by male, five by female speakers. In order to allow the subjects to get an impression of the residual noise by itself, the speech samples were preceded and followed by speech pauses of 3 s overall duration. The average duration of the resulting samples was about 7 s. The noise was scaled and added such that the noisy samples had an average segmental SNR of 0 dB in frames where speech is present. Each of the noisy samples was filtered by the conventional and proposed approach, respectively, resulting in ten pairs of enhanced samples per noise type. The participants were asked to select the file in each pair they preferred in terms of speech quality, naturalness of the background and overall quality, respectively. The comparison was done blindly and in randomized order. The participants were divided into two groups, experts and non-experts. While the 7 expert listeners clearly favored the proposed cepstral smoothing approach, we present detailed results only for the 12 non-expert listeners for babble and pink noise in Table 3.2. It may be seen that especially in nonstationary environments, the participants favored the cepstral approach. This is because the background noise sounds less tonal and thus more natural with the proposed approach. This is achieved without affecting the speech quality. On the contrary: for stationary noise sources, where both algorithms perform equally well in terms of background quality (no musical noise), a preference for TCS in terms of speech and overall quality may be seen. The audio examples are available at [Breithaupt and Gerkmann, 2007].

Noise	Category	TCS of spectral gain functions	decision-directed [Ephraim and Malah, 1984] with $\alpha_{\mathrm{dd}} = 0.97$ and [Malah *et al*, 1999]	Equally Suited
Babble	Backgr.	68%	5%	27%
	Speech	54%	8%	38%
	Overall	75%	7%	18%
Subway	Backgr.	61%	5%	34%
	Speech	69%	8%	23%
	Overall	84%	4%	12%
Pink	Backgr.	18%	18%	64%
	Speech	52%	23%	25%
	Overall	50%	22%	28%

Table 3.2: The results of the listening experiment for babble, subway, and pink noise. The numbers state the percentage of votes in favor of one of the filters. The choice "equally suited" was also possible.

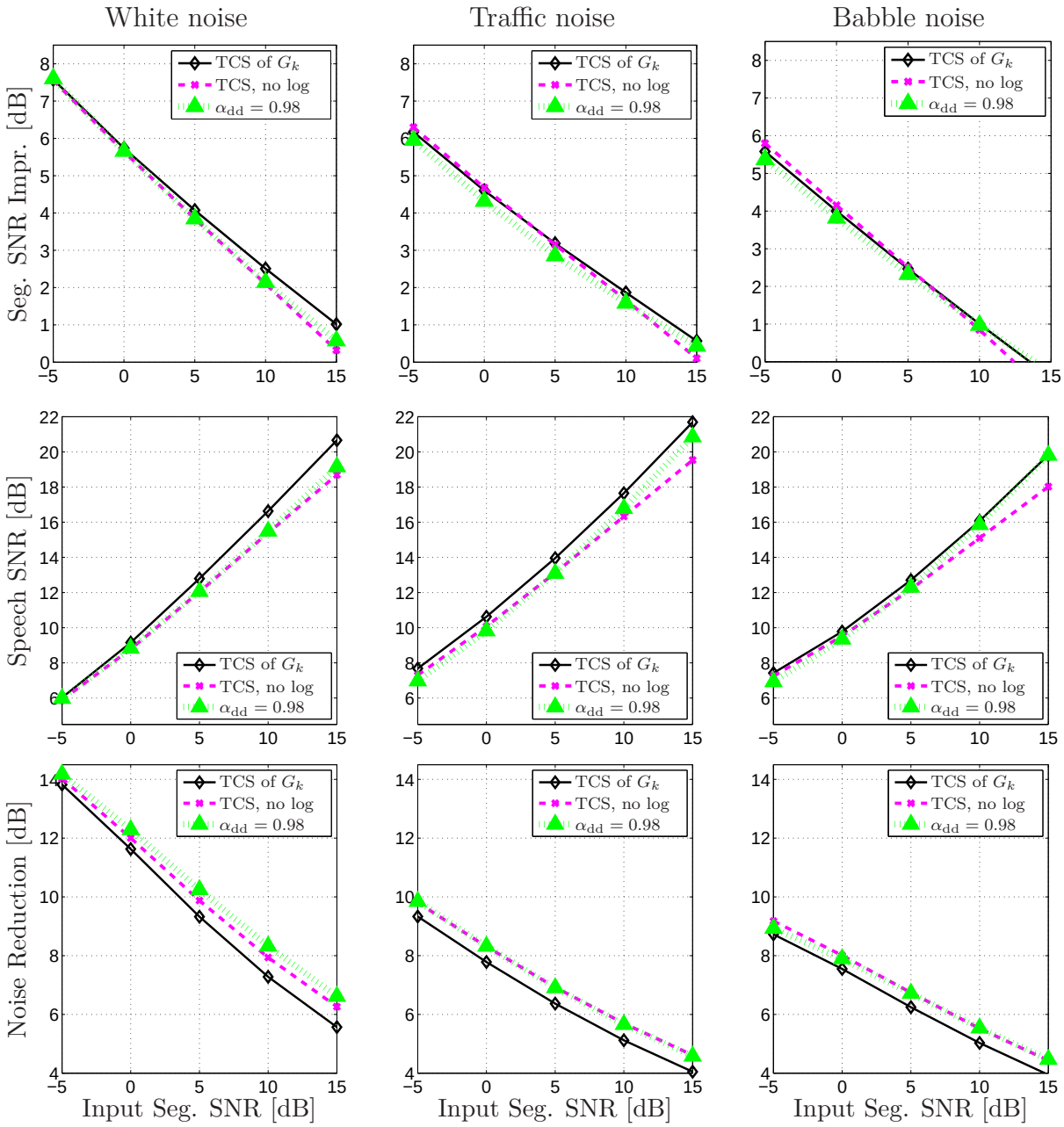


Figure 3.3: Instrumental measures for a TCS of spectral gain functions and the decision-directed approach with $\alpha_{\mathrm{dd}} = 0.98$.

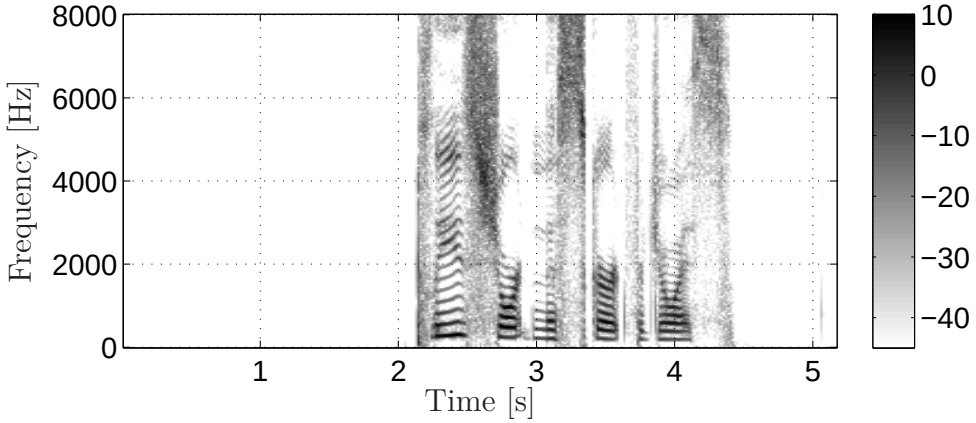
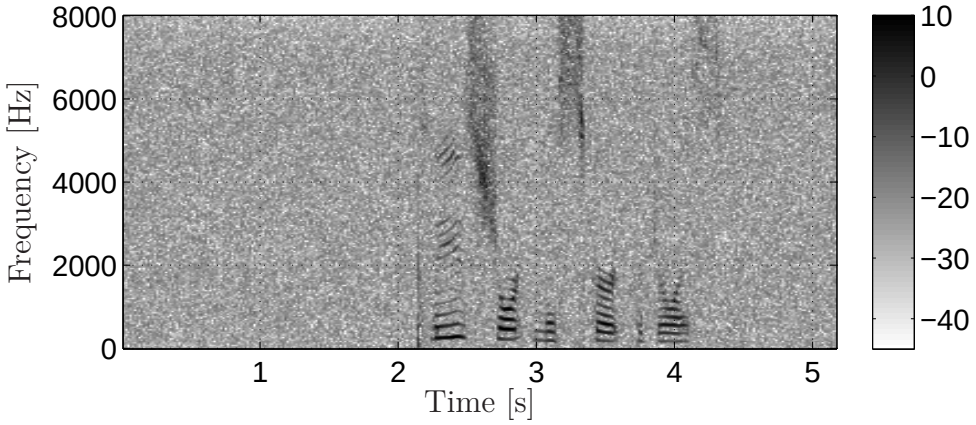
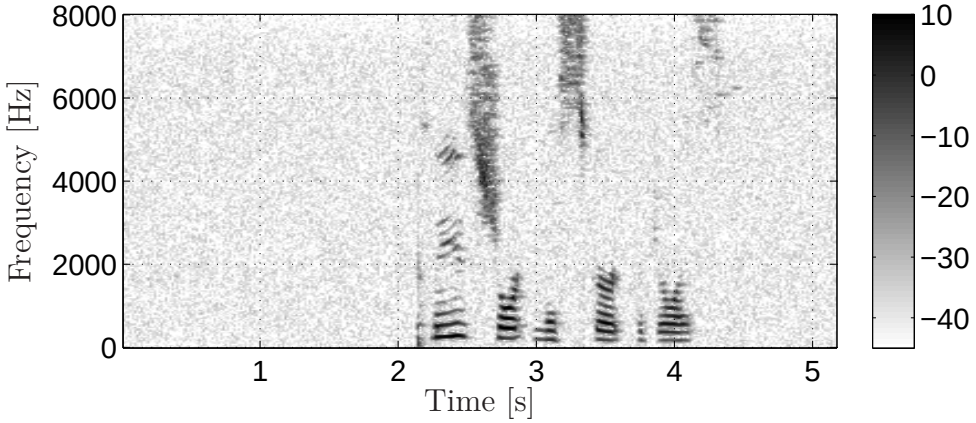


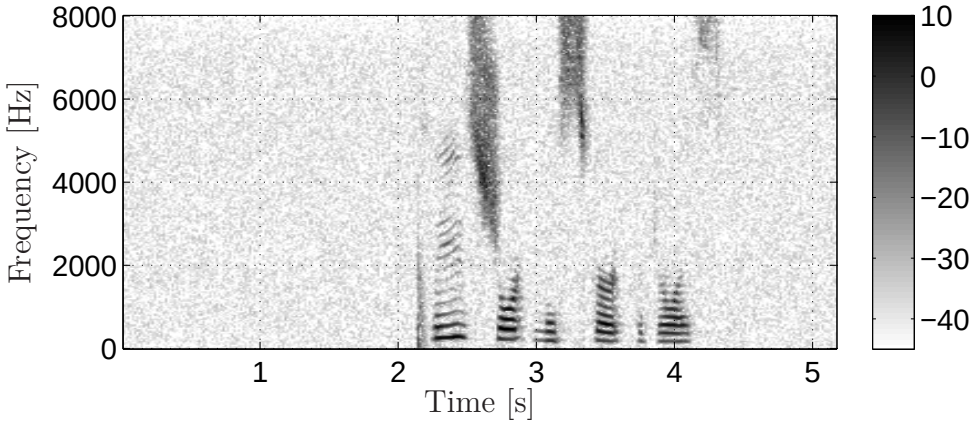
Figure 3.4: Spectrogram for the sentence “Please shorten this skirt for Joice.” spoken by a female.



(a) Noisy Speech

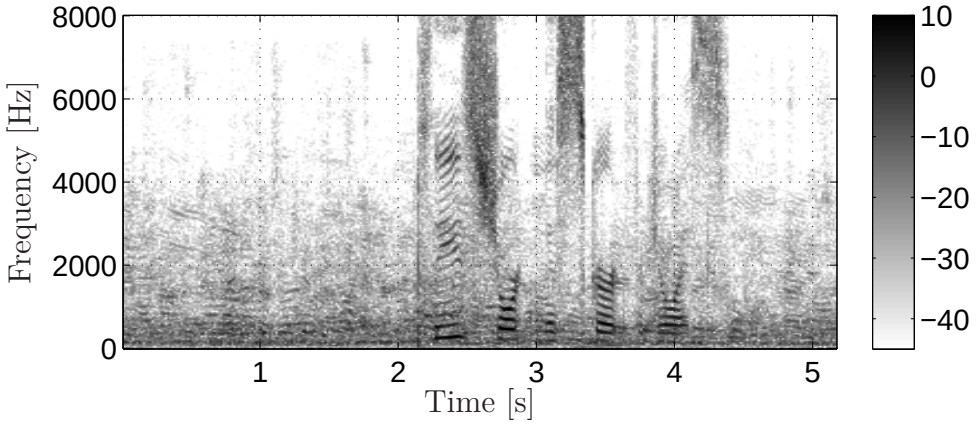


(b) Enhanced speech with the decision-directed approach and $\alpha_{\mathrm{dd}} = 0.98$

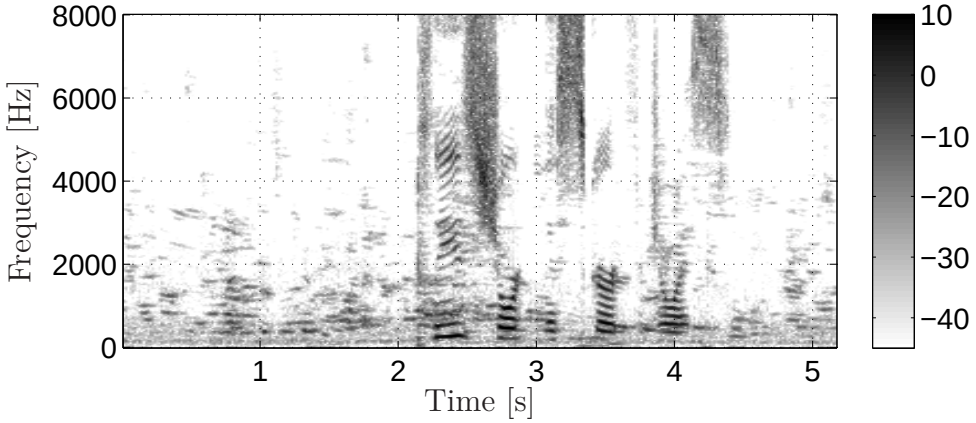


(c) Cepstral Smoothing of spectral gain functions

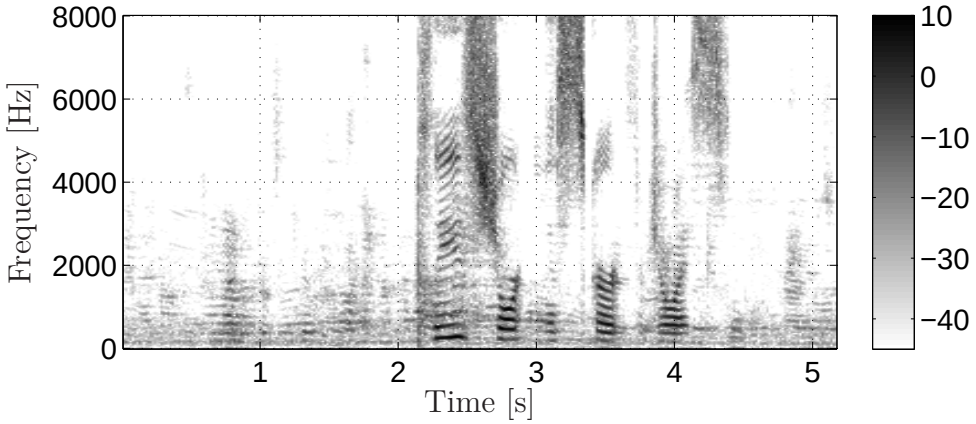
Figure 3.5: Spectrograms for noisy and enhanced speech. The clean speech of Figure 3.4 is disturbed by stationary white Gaussian noise at 0 dB segmental SNR.



(a) Noisy Speech



(b) Enhanced speech with the decision-directed approach and $\alpha_{\mathrm{dd}} = 0.98$



(c) Cepstral Smoothing of spectral gain functions

Figure 3.6: Spectrograms for noisy and enhanced speech. The clean speech of Figure 3.4 is disturbed by nonstationary babble noise at 0 dB segmental SNR.

3.2 Temporal cepstrum smoothing for *a priori* SNR estimation

While a smoothing of the spectral gain function is a very general tool and can be applied to any speech enhancement algorithm where the output is achieved with a multiplicative spectral weighting, an even better performance may be expected if TCS is applied in an earlier step of the speech enhancement algorithm. If the spectral gain function is given by the Wiener Filter

$$G_k = \frac{\xi_k}{1+\xi_k}\,,$$

the next earlier step is the estimation of the *a priori* SNR ξ_k. In this section we propose to use TCS to replace the well known *decision-directed a priori* SNR estimator [Ephraim and Malah, 1984].

3.2.1 Revision of *a priori* SNR estimation

The estimation of the *a priori* SNR is a very important part in speech enhancement algorithms. Erroneous estimation of this parameter leads either to a reduced noise reduction, speech distortions or musical noise. In nonstationary noise environments the estimation of the *a priori* SNR is particularly difficult. In this section we review state-of-the art *a priori* SNR estimators, particularly the well known decision-directed approach introduced by [Ephraim and Malah, 1984], and discuss their strengths and weaknesses. We then propose to estimate the *a priori* SNR by temporally smoothing the Maximum Likelihood (ML) *a priori* SNR estimate in the cepstral domain.

The *a priori* SNR, ξ_k, is defined as the ratio of the speech power, $\sigma_{\mathrm{S},k}^2$, and the noise power, $\sigma_{\mathrm{N},k}^2$. In Appendix B.1 we show that for a χ^2-distributed *a posteriori* SNR $\gamma_k = |Y_k|^2/\sigma_{\mathrm{N},k}^2$ the ML estimate of the *a priori* SNR is given as

$$\xi_k^{\mathrm{ml}} = \gamma_k - 1 = \frac{|S_k + N_k|^2 - \mathrm{E}\{|N_k|^2\}}{\mathrm{E}\{|N_k|^2\}}\,. \tag{3.12}$$

If speech and noise are uncorrelated, the expected value of the ML estimate is the *a priori* SNR:

$$\xi_k = \frac{\sigma_{\mathrm{S},k}^2}{\sigma_{\mathrm{N},k}^2} = \frac{\mathrm{E}\{|S_k|^2\} + \mathrm{E}\{|N_k|^2\} - \mathrm{E}\{|N_k|^2\}}{\mathrm{E}\{|N_k|^2\}} = \mathrm{E}\Big\{\xi_k^{\mathrm{ml}}\Big\}\,. \tag{3.13}$$

Thus, the ML *a priori* SNR ξ_k^{ml} is an unbiased estimate of the *a priori* SNR ξ_k. However, any deviation of $|N_k|^2$ from the noise power $\sigma_{\mathrm{N},k}^2 = \mathrm{E}\{|N_k|^2\}$ will cause fluctuations in the ML SNR estimator, ξ_k^{ml}. When employed in a speech enhancement framework, *e.g.* for computing the Wiener filter (1.4), these fluctuations yield a very unnatural sounding residual noise. In [Ephraim and Malah, 1984], before introducing the decision-directed approach, Ephraim and Malah derived an ML estimator based on consecutive analysis frames that results in a recursive smoothing of (3.12). This recursive smoothing can be interpreted as an approximation of the true *a priori* SNR $\xi_k = \mathrm{E}\left\{\xi_k^{\mathrm{ml}}\right\}$, assuming that the speech signal is ergodic. However, since speech is highly non-stationary (and hence not ergodic), recursive smoothing results in a poor trade-off between fluctuations in the residual noise and distortions of speech onsets and transients. If the recursive smoothing constant is chosen high enough to eliminate fluctuations in ξ_k^{ml}, it also distorts speech onsets and transitions, resulting in a reduced speech quality. Therefore, in state-of-the-art speech enhancement algorithms the *a priori* SNR is estimated in a decision-directed way [Ephraim and Malah, 1984, Ephraim and Cohen, 2006], *i.e.* based on a previous clean-speech estimate $\widehat{S}_k(l-1)$ which may be obtained with (1.3) using the Wiener filter (1.4):

$$\widehat{\xi}_k(l) = \max\left\{\alpha_{\mathrm{dd}}\frac{|\widehat{S}_k(l-1)\,|^2}{\sigma_{\mathrm{N},k}^2(l-1)} + (1-\alpha_{\mathrm{dd}})\xi_k^{\mathrm{ml}}(l)\,,\xi_{\min}\right\}. \tag{3.14}$$

The parameters α_{dd} and $\xi_{\min}$ control the trade-off between noise reduction and distortions of speech transients in a speech enhancement framework [Cappé, 1994]. The decision-directed procedure (3.14) allows for a fast tracking of increasing levels of the speech power, thus effectively resulting in an adaptive smoothing. Consequently, at speech onsets and transitions, less speech distortions are introduced as compared to a recursive smoothing of ξ_k^{ml}.

However, since the decision-directed SNR estimator is sensitive to rising spectral amplitudes, it does not only respond to speech onsets, but also to noise bursts that are not tracked by the noise power estimation algorithm. Therefore, noise bursts will cause a rising *a priori* SNR estimate, and thus outliers in the residual noise of the clean-speech estimate that may be perceived as annoying musical tones. Furthermore, the performance of the decision-directed approach depends on the type of speech estimator $G_{\mathcal{H}_{1,k},k}$ in (1.3) [Breithaupt and Martin, 2010]. The SNR estimation approach proposed next is capable of avoiding annoying outliers while preserving the speech characteristics. Further, it also decouples the estimation of the clean speech spectral coefficients and the *a priori* SNR estimation.

3.2.2 Proposed *a priori* SNR estimation

From the ML SNR estimate (3.12) we compute the speech power

$$\sigma_{\mathrm{S},k}^{2,\,\mathrm{ml}} = \sigma_{\mathrm{N},k}^{2}\, \max\left\{\xi_k^{\,\mathrm{ml}}, \xi_{\min}^{\,\mathrm{ml}}\right\}. \tag{3.15}$$

Here $\xi_{\min}^{\,\mathrm{ml}} > 0$ is a small lower bound which prevents $\xi^{\,\mathrm{ml}}$ from taking negative values or values close to zero and thus avoids numerical difficulties in the following steps. We then apply TCS to the ML speech power estimate using Algorithm 1 where $\Phi_k = \sigma_{\mathrm{S},k}^{2,\,\mathrm{ml}}$. The smoothed output $\bar{\Phi}_k$ represents the speech power estimate $\hat{\sigma}_{\mathrm{S},k}^2$, such that the *a priori* SNR estimate is gained as

$$\hat{\xi}_k = \max\left\{\frac{\hat{\sigma}_{\mathrm{S},k}^2}{\sigma_{\mathrm{N},k}^2}, \xi_{\min}\right\}. \tag{3.16}$$

The bias correction $\mathcal{B}$ in Algorithm 1 is computed using Algorithm 4 where we assume $\sigma_{\mathrm{S},k}^{2,\,\mathrm{ml}}$ to be χ^2-distributed with $\mu = 1$. In Section 3.2.4 the proposed TCS for *a priori* SNR estimation is compared to the decision-directed *a priori* SNR estimator [Ephraim and Malah, 1984].

3.2.3 Reduction of the computational complexity

As both the spectral power and the cepstral coefficients are real and symmetric with respect to $N/2$, the N point DFT and its inverse in steps 2 and 6 of Algorithm 1 can also be computed by a type-I $N/2+1$ point Discrete Cosine Transform (DCT) [Wang, 1984, Wang, 1991].

$$\mathrm{DFT}\{\phi_q\} = \sum_{q=0}^{N/2} \vartheta_q \phi_q \cos\left(\frac{\pi q k}{N/2}\right) = \mathrm{DCT}\left\{\phi_q\right\}, \tag{3.17}$$

where $\vartheta_q = 1$ for $q \in \{0, N/2\}$ and $\vartheta_q = 2$ for $q \notin \{0, N/2\}$. The Inverse Discrete Cosine Transform (IDCT) results from a simple scaling of the DCT by $1/N$. While the general complex N point DFT exhibits a computational complexity of the order $N\log_2(N)$ if realized in terms of a Fast Fourier Transform (FFT) [Cooley and Tukey, 1965], a fast DCT on half the length exhibits a complexity of the order $N/4\log_2(N/2)$ [Wang, 1991, Lo and Cham, 1996].

To additionally reduce the computational complexity we propose to replace the strong smoothing of the upper cepstral coefficients by cepstral nulling. If a subset of only $L < N/2+1$ connected cepstral coefficients is assumed nonzero, a *pruned* DCT can

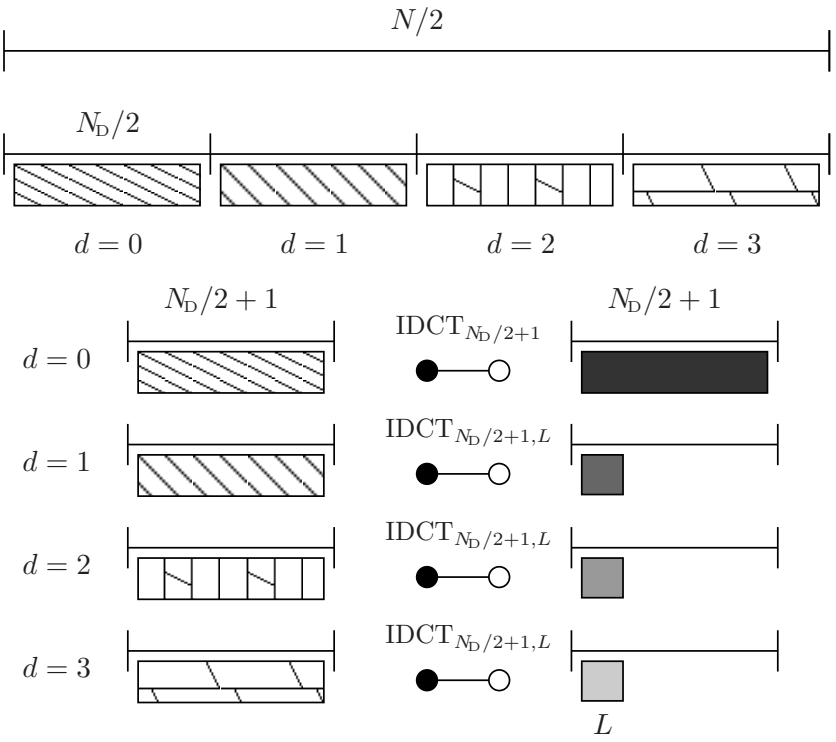


Figure 3.7: Illustration of TCS with reduced computational complexity with $D = 4$. While $\mathrm{IDCT}_{N_\mathrm{D}/2+1}$ denotes an $N_\mathrm{D}/2+1$ point IDCT, $\mathrm{IDCT}_{N_\mathrm{D}/2+1,L}$ denotes a pruned $N_\mathrm{D}/2+1$ point IDCT with pruned length L.

be used instead of a regular $N/2+1$ point DCT which reduces the computational complexity approximately by the factor $\log_2(L)\,/\,\log_2(N/2)$ [Wang, 1991]. However, in speech processing it is important that the fundamental period peak in the cepstral representation is preserved. As the position of the speech fundamental period peak may lie anywhere in the range $q \in \{f_\mathrm{s}/800\,\mathrm{Hz}, ..., N/2\}$, a pruned DCT with a fixed pruned length L is not applicable. The spectral harmonics of the speech fundamental frequency are especially strong and important in the low frequency range, *e.g.* below 2 kHz. Hence, we divide the spectrum into a low frequency range, where the spectral harmonics are preserved, and the high frequencies, where a pruned DCT is used. To keep the DCT lengths of the low and high frequency divisions as powers of two, we divide the spectrum into D subdivisions, where D is also a power of two. Thus, we now have D subdivisions indexed by d, with a bandwidth of $f_\mathrm{s}/(2D)$ and a segment length of $N_\mathrm{D}/2+1 = N/(2D)+1$ each (cf. Figure 3.7), where

$$N_\mathrm{D} = N/D. \tag{3.18}$$

The exact arrangement of the subdivisions in the frequency domain (left hand side of

Figure 3.7) is given in (3.19), where Φ_k is short for $\sigma_{\mathrm{S},k}^{2,\mathrm{ml}}$.

$$\begin{aligned} d=0: &\quad \Big[\Phi_0, \quad \Phi_1, \quad ..., \quad \Phi_{\frac{N_\mathrm{D}}{2}-1}, \quad 0\Big] \\ d>0: &\quad \Big[0, \quad \Phi_{d\frac{N_\mathrm{D}}{2}}, \quad ..., \quad \Phi_{(d+1)\frac{N_\mathrm{D}}{2}-2}, \quad \Phi_{(d+1)\frac{N_\mathrm{D}}{2}-1}\Big]. \end{aligned} \tag{3.19}$$

The Nyquist bin $\Phi_{N/2}$ is neglected such that the spectral coefficients for all subdivisions but the zeroth, obey the same distribution model. For resynthesis we keep the non-smoothed Nyquist bin $\Phi_{N/2}$.

While for the zeroth subdivision we apply a regular TCS, for the remaining subdivisions $d = 1, ..., D-1$ we use TCS for the lower cepstral coefficients and null the cepstral coefficients $q > L$. This allows us to use a pruned $N_\mathrm{D}/2+1$ point DCT for $D-1$ subdivisions with pruned length L. For $d > 0$, we find an effective length of the pruned data of 1 ms to be sufficient to represent the speech spectral envelope. This results in an effective length of the pruned data of

$$L = 1\,\mathrm{ms} \cdot f_\mathrm{s}/D = 16/D. \tag{3.20}$$

The relative computational complexity with respect to the $N/2+1$ point DCT is

$$\mathcal{C} = \frac{N_\mathrm{D}/4\log_2(N_\mathrm{D}/2) + N_\mathrm{D}/4(D-1)\log_2(L)}{N/4\log_2(N/2)}\,. \tag{3.21}$$

For $D = 2$ and $N = 512$ the proposed approach with $D = 2$ requires $\mathcal{C}_{D=2} = 62.5\%$ of the computational complexity as compared to a $N/2+1$ point DCT without pruning. For $D = 4$ we obtain $\mathcal{C}_{D=4} = 37.5\%$. When we compare the proposed approach to using a complex FFT on the full symmetric spectrum, the relative complexity is as low as $\mathcal{C}_{D=2} = 13.9\%$ and $\mathcal{C}_{D=4} = 8.3\%$, respectively.

For the bias correction we employ Algorithm 4, where now N_D is used instead of N in (2.21) and Algorithm 4. For the pruned subdivisions $d > 0$, the sum in step 4 of Algorithm 4 is only computed up to L instead of $N_\mathrm{D}/2$.

The results of Section 3.2.3 are accepted for publication in [Gerkmann and Martin, 2010a].

3.2.4 Experimental results

We now compare TCS for speech power estimation to the decision-directed approach for single channel speech enhancement. For the filter function we use the Wiener filter and assume that speech is present in all time-frequency points. Thus, the gain function in (1.3) is given by $G_k = \hat{\xi}_k/(1+\hat{\xi}_k)$, and then limited to be larger then G_min via (1.7).

The spectral noise power is estimated using the minimum statistics approach [Martin, 2001]. For the short-time Fourier analysis (1.1) we use Hann windows w_n with a length of 32 ms and 50% overlap.

For the recursive smoothing constant in (3.4), we choose:

$$\alpha_q^{\text{const}} = \begin{cases} 0.5 & , q \in \{0, ..., \lfloor 2/D \rfloor\} \\ 0.7 & , q \in \{\lfloor 2/D \rfloor + 1, ..., \lfloor 20/D \rfloor - 1\} \\ 0.97 & , q \in \{\lfloor 20/D \rfloor, ..., N/(2D)\}\,, \end{cases} \tag{3.22}$$

where $\lfloor \cdot \rfloor$ is the flooring operator. For the subdivisions $d > 0$ we only smooth the first $L = 16/D$ bins and $\mathbb{Q}_{\text{pitch}}$ is an empty set, no matter if the speech segment is voiced or unvoiced. The remaining cepstral coefficients $q \geq L$ are implicitly set to zero by applying pruned DCTs for the cepstral transform and its inverse. The remaining parameters are summarized in Table 3.3. Apart from the bias correction and $\xi_{\min}^{\text{ml}}$, the used parameters are identical to [Breithaupt *et al*, 2008a] for $D = 1$.

Smoothing factor for $\mathbb{Q}_{\text{pitch}}$ (3.4)	α_{pitch}	$= 0.2$
Threshold for voiced/unvoiced decision (3.3)	Λ^{thr}	$= 0.2 \cdot D$
Lower bound for the q_0 search (3.2)	q_{low}	$= \left\lfloor \frac{f_{\text{s}}}{D \cdot 300\,\text{Hz}} \right\rfloor$
Upper bound for the q_0 search (3.2)	q_{high}	$= \left\lfloor \frac{f_{\text{s}}}{D \cdot 70\,\text{Hz}} \right\rfloor$
Margin for $\mathbb{Q}_{\text{pitch}}$ (3.3)	Δq_0	$= \lfloor 2/D \rfloor$
Length of the cepstral low-pass (3.1)	τ_{H}	$= \lfloor \frac{f_{\text{s}}}{D \cdot 2000\,\text{Hz}} \rfloor$
Smoothing constant for (3.4)	β	$= 0.96$
Lower bound on the *a priori* SNR (3.16)	$10\log_{10}(\xi_{\min})$	$= -25\,\text{dB}$
Lower bound on the *a priori* SNR (3.15)	$10\log_{10}(\xi_{\min}^{\text{ml}})$	$= -30\,\text{dB}$
Lower bound on the gain function (1.7)	$20\log_{10}(G_{\min})$	$= -17\,\text{dB}$
Sampling rate	f_{s}	$= 16\,\text{kHz}$

Table 3.3: Parameters for TCS based spectral *a priori* SNR estimation.

As in Section 3.1, we process 320 speech samples of [Garofolo, 1988, dialect region 6] that are disturbed by several noise types and input SNRs. Again, we evaluate the algorithms in terms of the segmental SNR improvement, the segmental speech SNR and noise reduction [Breithaupt, 2008, Lotter, 2004, Lotter and Vary, 2005]. The results are given in Figure 3.8. We first compare the TCS approach with $D = 1$ to the

decision-directed approach (3.14), where we set the smoothing constant to $\alpha_{\mathrm{dd}} = 0.98$, as proposed in [Ephraim and Malah, 1984]. It is obvious, that the TCS approach outperforms the decision-directed approach [Ephraim and Malah, 1984] in terms of the segmental SNR improvement and the speech SNR while performing virtually the same in terms of noise reduction. The performance is also better as compared to a TCS of spectral gain functions given in Figure 3.3. This can be attributed to the fact that TCS for *a priori* SNR estimation is used in an earlier step of the enhancement framework. The advantage of a TCS of the spectral gain function is that it can be more flexibly used in any speech enhancement algorithm that uses multiplicative spectral gain functions, including binary masks as used in blind source separation, *e.g.* [Jan *et al*, 2009].

In figures 3.9 and 3.10, the spectrograms of enhanced speech using the TCS based *a priori* SNR estimator are given. As compared to the results of the decision-directed approach given in figures 3.5 and 3.6, it can be observed, that for babble noise a much smoother background noise can be observed. Informal listening confirms that the TCS based approach results in a much more natural sounding residual noise with a strong reduction of musical tones, and also a better speech quality as compared to the decision-directed approach. Especially in the case of white Gaussian noise the processed speech using TCS sounds clearer, as more low-energy speech components are preserved. As for the residual noise, in the case of white noise, neither approach produces musical noise.

In Figure 3.8 also the results with reduced complexity are given, *i.e.* $D = 2$ and $D = 4$. It can be seen that the proposed low complexity approach with two subdivisions exhibits virtually the same performance as the reference method $D = 1$ with computational savings of 37.5%. The method with four subdivisions exhibits computational savings of 62.5% with respect to $D = 1$ while still performing considerably better than the decision-directed approach [Ephraim and Malah, 1984]. For babble noise, the approach with $D = 4$ results in a larger noise reduction than for $D = 2$ and $D = 1$. However, compared to $D = 1$ and $D = 2$ the speech SNR and the segmental SNR are generally reduced when four subdivisions are used. The reason why the approach with two subdivisions performs better than the approach with four subdivisions is that in the first case the spectral harmonics of the speech power estimate are preserved up to a frequency of 4 kHz while in the latter the spectral harmonics are only preserved up to a frequency of 2 kHz.

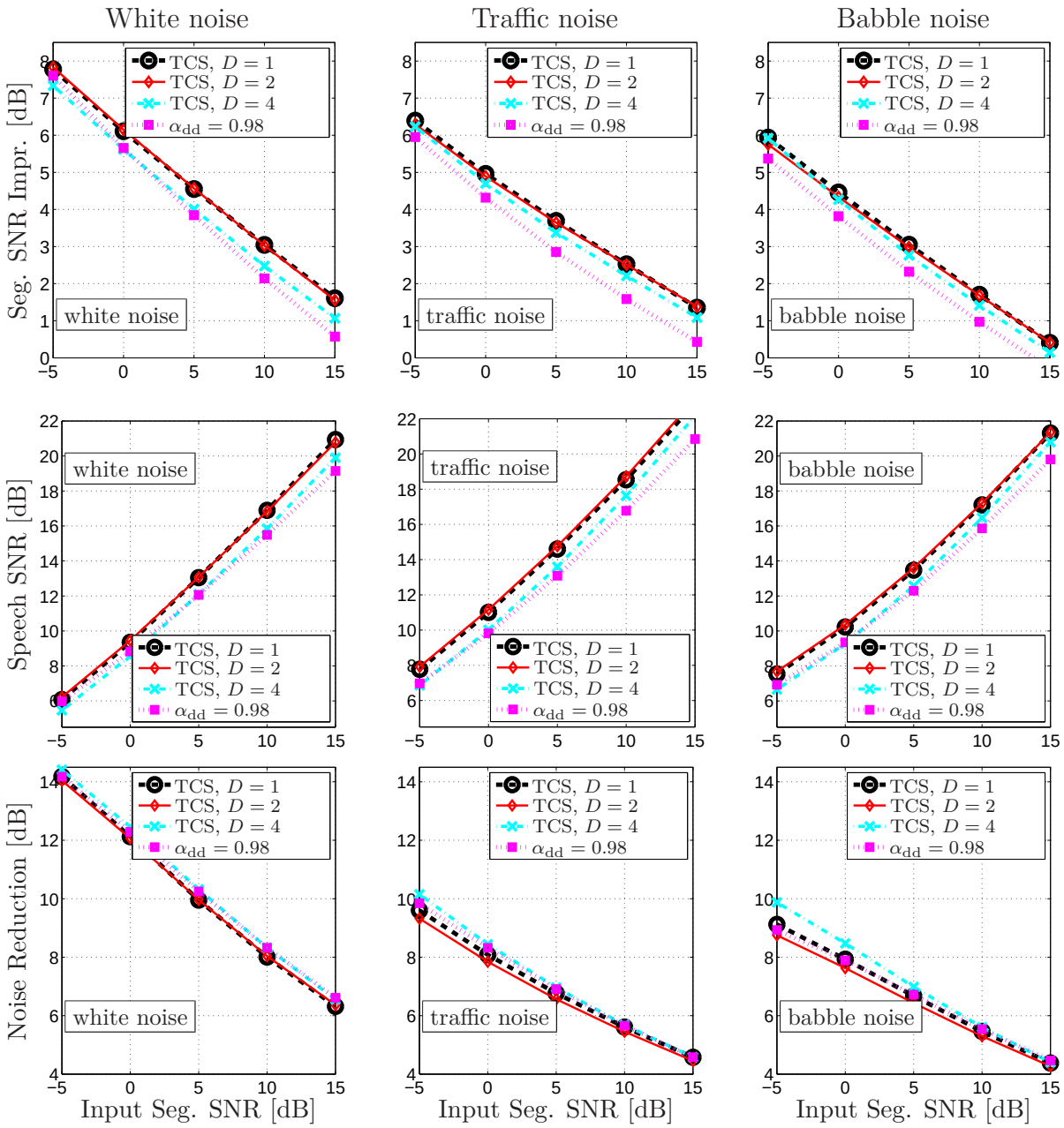


Figure 3.8: Instrumental measures for TCS based *a priori* SNR estimation and the decision-directed approach with $\alpha_{\mathrm{dd}} = 0.98$.

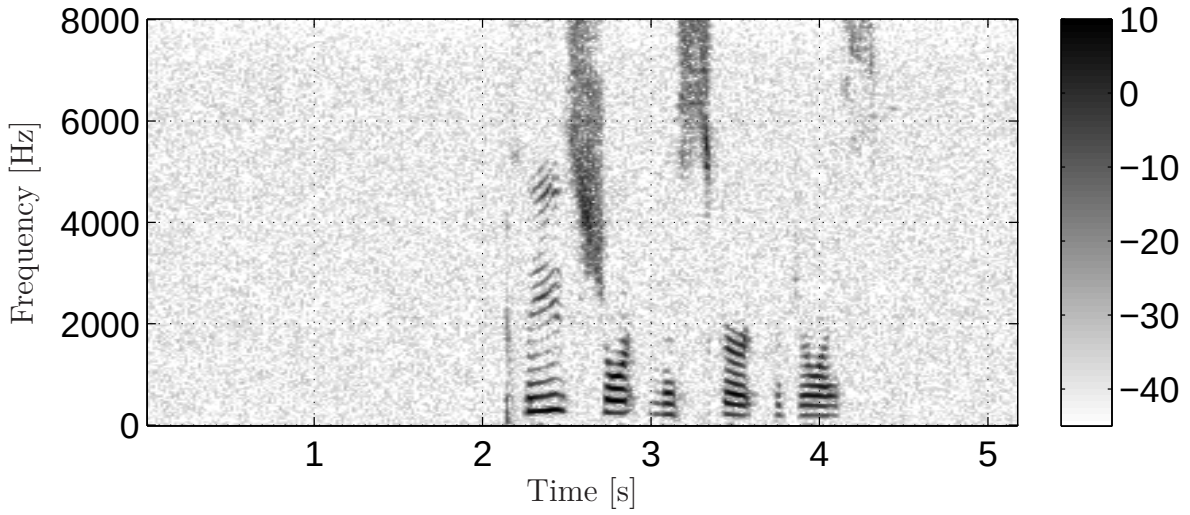


Figure 3.9: Spectrogram of enhanced speech using TCS for *a priori* SNR estimation with $D = 1$. The clean speech of Figure 3.4 is disturbed by stationary white Gaussian noise at 0 dB segmental SNR. The noisy speech is given in Figure 3.5.

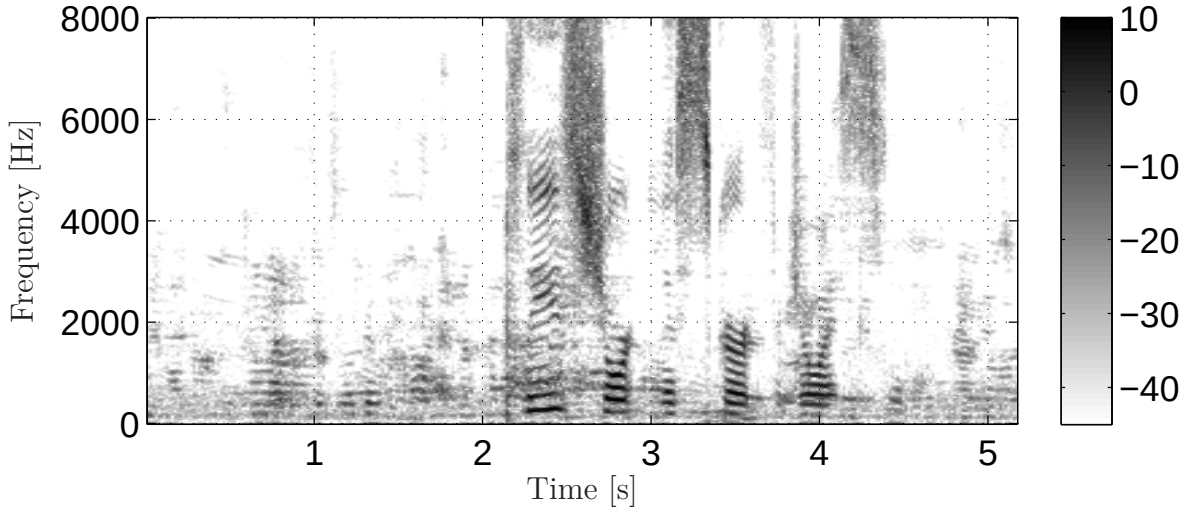


Figure 3.10: Spectrogram of enhanced speech using TCS for *a priori* SNR estimation with $D = 1$. The clean speech of Figure 3.4 is disturbed by nonstationary babble noise at 0 dB segmental SNR. The noisy speech is given in Figure 3.6.

3.3 Conclusions

In this chapter, the concept of Temporal Cepstrum Smoothing (TCS) is incorporated to different parts of single channel speech enhancement algorithms. The first approach, discussed in Section 3.1, applies TCS to the spectral gain function. Listening experiments have indicated that this results in a more natural sounding residual noise and a higher signal quality. In the second approach TCS is used to estimate the *a priori* Signal-to-Noise Ratio (SNR) in speech enhancement algorithms, which produces even better results in terms of instrumental measures as compared to a TCS of spectral gain functions. This can be attributed to the fact that TCS is incorporated in an earlier step of the noise reduction algorithm. The advantage of smoothing spectral gain functions is that it can be employed to reduce spectral outliers in any speech enhancement algorithm that estimates clean speech spectral coefficients by applying a multiplicative gain function or binary masks. The additional costs of cepstral smoothing approaches are dominated by the transformations needed to compute the cepstrum and its inverse. These transformations can be efficiently computed using pruned fast Discrete Cosine Transforms (DCTs).

Chapter 4

Instantaneous Cepstral Replacement Techniques

In this chapter, we modify estimated spectral quantities in the earliest possible step of the speech enhancement framework introduced in Chapter 1, with the aim that spectral outliers are not reduced, but *avoided* before they even occur. Furthermore, while a temporal smoothing of cepstral coefficients may smear the background noise over time, the approach proposed in this chapter reduces spectral outliers instantaneously. Especially in nonstationary noise, such as babble noise, this may yield a more natural sounding residual noise.

Spectral outliers occur, if the spectral noise power is locally underestimated, for instance because a babble burst of short duration is not tracked by the minimum statistics noise power estimator [Martin, 2001]. As a consequence, spectral outliers occur in the Maximum Likelihood (ML) *a priori* Signal-to-Noise Ratio (SNR) estimate that cannot be fully suppressed if the decision-directed approach (3.14) is used for the estimation of the *a priori* SNR [Breithaupt and Martin, 2010] but are reduced if Temporal Cepstrum Smoothing (TCS) is used for *a priori* SNR estimation (Section 3.2). In this chapter we modify the estimate of the spectral noise power, so that spectral outliers in the *a priori* SNR estimate are avoided before they occur. However, a modification of the spectral noise power estimate has to be done with great care: if the spectral noise power is too large, this results in an underestimation of the *a priori* SNR and in speech distortions when applied in a speech enhancement framework. Therefore, in this chapter we combine a careful modification of the spectral noise power with a modification of the spectral speech power.

4.1 Cepstral modification of the spectral noise power

The minimum statistics approach [Martin, 2001] is a well established algorithm for spectral noise power estimation. The basic assumption of the minimum statistics approach is that the noise signal is more stationary than the speech signal. It is assumed that in each frequency bin within a 1.5 second window speech is not active in at least one segment l, such that the minimum of the smoothed squared noisy observation in the 1.5 second window can be attributed to the noise signal only. A bias correction is necessary to infer the mean of the smoothed squared noise spectral coefficients from the found minimum. The inferred mean represents the minimum statistics spectral noise power estimate. One of the most powerful aspects of this approach is that it results in very little speech distortions. However, noise bursts of short duration cannot be tracked by the minimum statistics spectral noise power estimator and are likely to result in musical noise.

In this section, we propose to replace the cepstral coefficients $q \in \overline{\mathbb{Q}}$ (defined in (2.3)) of the cepstral representation of the spectral noise power by the corresponding coefficients of the cepstral representation of the magnitude squared noisy observation. As a result, by exploiting the *a priori* knowledge that the speech spectral coefficients are reflected mostly by the cepstral coefficients $q \in \mathbb{Q}$, the modified spectral noise power estimate also contains the spectral fine structure of the non speech related cepstral coefficients $q \in \overline{\mathbb{Q}}$ of the noisy observation.

However, when replacing coefficients of the noise power by coefficients of the noisy observation, one has to be very careful that no speech information leaks into the noise power estimate, as this may result in speech distortions when applied in a speech enhancement framework. Thus, the lower cepstral bound q_{low} in (2.3) should be carefully chosen: if it is too large, musical noise cannot be effectively suppressed, if it is too low, speech distortions may occur. In Section 2.1, we analyzed which cepstral coefficients carry the most information about speech. As a result, it turned out that more cepstral coefficients are needed to represent the spectral envelope of voiced speech sounds than for unvoiced speech sounds. As a consequence, for a cepstral modification of the spectral noise power, the lower bound for voiced sounds $q_{\text{low,v}}$ should be chosen larger than the lower bound for unvoiced sounds $q_{\text{low,uv}}$, to avoid musical noise without introducing speech distortions.

4.1.1 Bias compensation

The spectral noise power estimate gained by using the minimum statistics approach is based on an optimal smoothing of the magnitude squared spectral noisy observation

[Martin and Lotter, 2001]. Assuming that the magnitude squared noisy observation is χ^2-distributed, the optimal unbiased smoothing reduces the variance of the noisy observation, but also increases the shape parameter $\mu_{\text{MS}} = (\text{E}\{|Y_k|^2\})^2/\text{var}\{|Y_{\text{MS},k}|^2\}$, where we denote the quantities after the smoothing of the minimum statistics by the index MS. To enable an instantaneous adaption to the fine structure of the noise signal, we propose to replace the cepstral coefficients $q \in \overline{\mathbb{Q}}$ by the instantaneous nonsmoothed noisy observation $|Y_k|^2$, which is assumed to be χ^2-distributed with $\mu = 1$. Thus, the resulting shape parameter $\bar{\mu}$ after the cepstral replacement will be reduced as compared to the shape parameter μ_{MS} before the replacement. This reduction of the shape parameter results in an overestimate of the spectral noise power, which can be compensated using a variation of Algorithm 4. The bias (2.34) now takes the form

$$\mathcal{B} = \frac{\mu_{\text{MS}}}{\bar{\mu}} \exp(\psi(\bar{\mu}) - \psi(\mu_{\text{MS}})) . \tag{4.1}$$

While the shape parameter μ_{MS} after spectral smoothing is estimated within the minimum statistics framework [Martin, 2001], the shape parameter after cepstral replacement $\bar{\mu}$ is estimated by constructing the variance after cepstral replacement via (2.21), where we choose the shape parameter to be μ_{MS} for the noise cepstral coefficients and $\mu = 1$ for those cepstral coefficients that have been replaced. $\bar{\mu}$ is then obtained by taking the sum of the variance of the cepstral coefficients after cepstral replacement, similar to (2.28).

4.2 Cepstral modification of the spectral speech power

For voiced sounds, cepstral coefficients of the noise power are only replaced for cepstral coefficients $q > q_{\text{low,v}}$ that represent the spectral fine structure. However, *e.g.* for babble noise, spectral noise bursts may be spectrally rather coarse such that a replacement of the cepstral coefficients $q_{\text{low,uv}} < q \leq q_{\text{low,v}}$ may substantially improve the elimination of processing artifacts. However, as mentioned above, replacing the coefficients $q_{\text{low,uv}} < q \leq q_{\text{low,v}}$ of the noise power estimate may yield a distortion of the spectral envelope of voiced sounds.

Alternatively to replacing cepstral coefficients of the noise power estimate, one may also replace coefficients of the speech power, estimated using *e.g.* the decision-directed approach (3.14). Spectral outliers due to processing artifacts result in a change of the spectral shape usually described by the cepstral coefficients $q \in \overline{\mathbb{Q}}$. Due to the modification of the noise power described in Section 4.1, for voiced sounds spectral outliers in the speech power estimate may still be expected to be represented by the cepstral coefficients $q_{\text{low,uv}} < q \leq q_{\text{low,v}}$. We thus propose to replace the cepstral coefficients

$q_{\text{low,uv}} < q \leq q_{\text{low,v}}$ of the speech power estimate by the corresponding cepstral coefficients of the cepstral representation of the magnitude squared noisy observation. As the spectral shape of the speech power estimate falls back to the spectral shape of the noisy observation, the proposed cepstral replacement of the speech power can be expected to result in considerably less speech distortions than a replacement in the noise power estimate, or a cepstral nulling of the speech power estimate.

The overall algorithm for the instantaneous cepstral replacement is given in Algorithm 5.

4.3 Evaluation

From the experiments in Section 2.1 we choose $q_{\text{low,v}} = 64$, which corresponds to $q_{\text{low,v}}/f_{\text{s}} = 4\,\text{ms}$, and $q_{\text{low,uv}} = 24$, which corresponds to $q_{\text{low,uv}}/f_{\text{s}} = 1.5\,\text{ms}$, assuming a sampling rate of $f_{\text{s}} = 16\,\text{kHz}$. The cepstral coefficients that represent the fundamental period of voiced sounds $q \in \mathbb{Q}_{\text{pitch}}$ are obtained using (3.2) and (3.3). To minimize speech distortions, for the cepstral modification of the noise power $\mathbb{Q}_{\text{pitch}}$ also includes $R = 3$ rahmonics rq_0 with $r = \{2, \ldots, R+1\}$ when $rq_0 \leq N/2$. If $\mathbb{Q}_{\text{pitch}}$ is the empty set, the signal segment l is assumed to be unvoiced and $q_{\text{low}} = q_{\text{low,uv}}$ in (2.2). If $\mathbb{Q}_{\text{pitch}}$ is not the empty set, the signal segment l is assumed voiced and $q_{\text{low}} = q_{\text{low,v}}$ in (2.2). We set the threshold $\Lambda^{\text{thr}} = 0.2$ and $\Delta q_0 = 2$ in (3.3). For the smoothing constant of the decision-directed approach α_{dd} in (3.14) we use $\alpha_{\text{dd}} = 0.94$ for the approach that uses cepstral replacement and $\alpha_{\text{dd}} = 0.98$ for the competing approach that does not use any cepstral techniques. For the short-time Fourier analysis (1.1) we use Hann windows w_n with a length of 32 ms and 50% overlap.

The results for 320 sentences from [Garofolo, 1988] are given in Figure 4.1. It can be seen that the performance is similar to the approach described in Section 3.2, where an ML estimate of the spectral speech power is temporally smoothed in the cepstral domain. In babble noise the proposed approach yields a slightly higher noise reduction for low SNRs which is also confirmed by informal listening. For traffic noise the segmental SNR is slightly lower for the cepstral replacement approach than for the TCS approach. This effect is almost inaudible though. Comparing Figure 4.1 to Figure 3.3, it can be seen that the instantaneous cepstral replacement outperforms a TCS of spectral gain functions in terms of instrumental measures. The spectrograms of enhanced speech using the instantaneous cepstral replacement are given in figures 4.2 and 4.3. As for all cepstral approaches presented in this thesis, for babble noise the residual background noise exhibits considerably less spectral outliers as compared to the decision-directed approach shown in Figure 3.6(b), while the speech signal is well preserved. As a result the amount of musical noise is reduced, while at the same time the speech SNR is higher

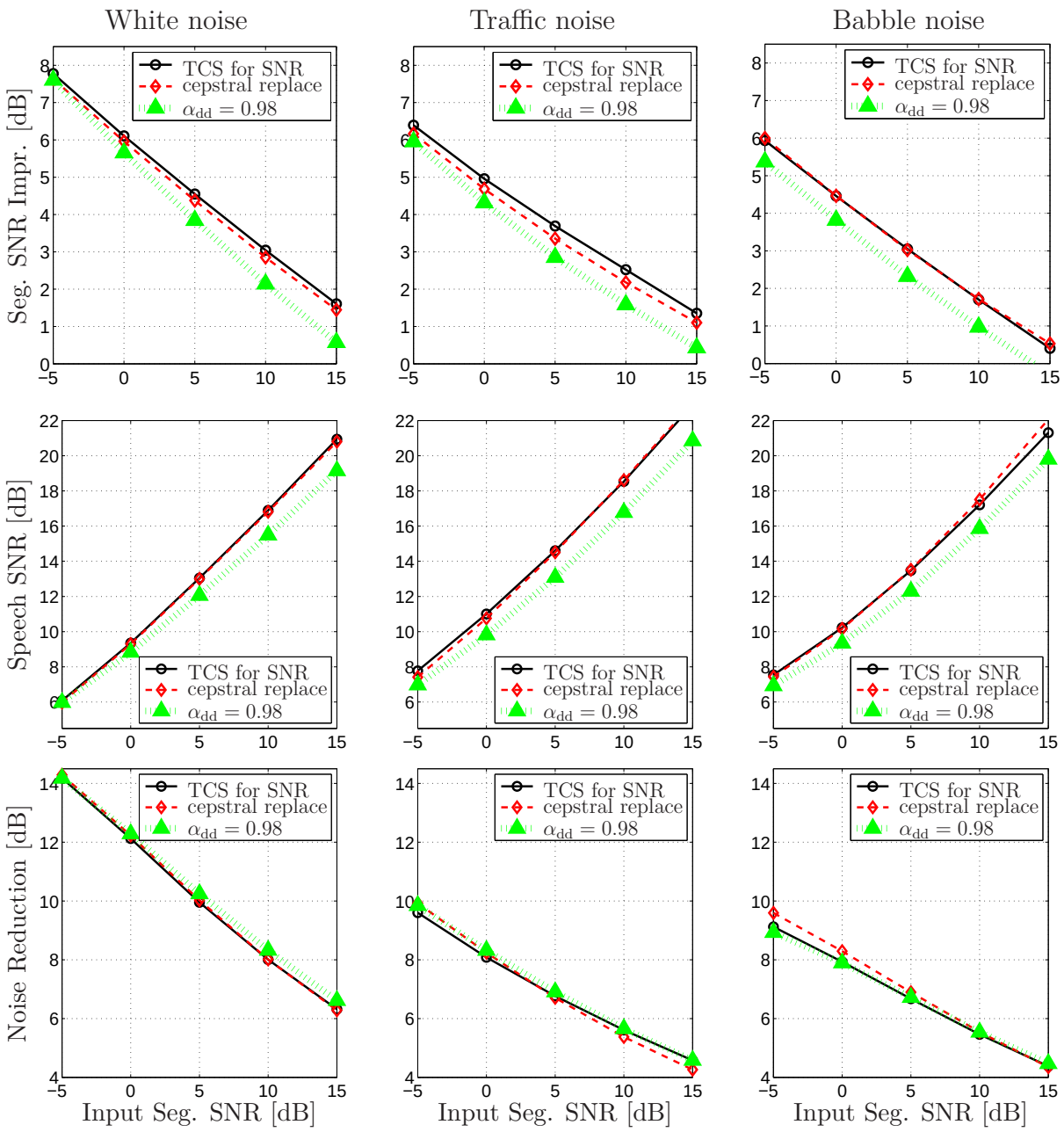


Figure 4.1: Results of instrumental measures for an instantaneous cepstral replacement in the noise and speech power estimates, TCS for *a priori* SNR estimation and the decision-directed approach with $\alpha_{dd} = 0.98$.

as compared to using the decision-directed approach with a larger smoothing factor α_{dd}.

Algorithm 5 Cepstral replacement of the speech and noise power estimates.

for all signal segments l **do**

Estimate the spectral noise power $\widehat{\sigma}^2_{\mathrm{N},k}$ using [Martin, 2001].

Cepstral transform of the spectral noise power and the noisy observation

$$\begin{aligned} \phi_{\mathrm{N},q} &= 1/N \sum_{k=0}^{N-1} \log\left(\widehat{\sigma}^2_{\mathrm{N},k}\right) \mathrm{e}^{\mathrm{j}2\pi kq/N}, \\ y_q &= 1/N \sum_{k=0}^{N-1} \log\left(|Y_k|^2\right) \mathrm{e}^{\mathrm{j}2\pi kq/N}. \end{aligned}$$

if segment l is voiced **then**

Moderate cepstral replacement

$$\bar{\phi}_{\mathrm{N},q} = \begin{cases} y_q & , q \in \{\{q_{\mathrm{low,v}}+1,\ldots,N/2\} \setminus \mathbb{Q}_{\mathrm{pitch}}\} \\ \phi_{\mathrm{N},q} & , \text{else.} \end{cases}$$

else

Cepstral replacement

$$\bar{\phi}_{\mathrm{N},q} = \begin{cases} y_q & , q \in \{q_{\mathrm{low,uv}}+1,\ldots,N/2\} \\ \phi_{\mathrm{N},q} & , \text{else.} \end{cases}$$

end if

Inverse cepstral transform (2.7) using (4.1)

$$\bar{\sigma}^2_{\mathrm{N},k} = \mathcal{B} \cdot \exp\left(\sum_{q=0}^{N-1} \bar{\phi}_{\mathrm{N},q}\, \mathrm{e}^{-\mathrm{j}2\pi kq/N}\right).$$

Estimate the spectral speech power $\widehat{\sigma}^2_{\mathrm{S},k}$ based on the modified noise power $\bar{\sigma}^2_{\mathrm{N},k}$ and the decision-directed approach (3.14).

Cepstral transform of the spectral speech power

$$\phi_{\mathrm{S},q} = 1/N \sum_{k=0}^{N-1} \log\left(\widehat{\sigma}^2_{\mathrm{S},k}\right) \mathrm{e}^{\mathrm{j}2\pi kq/N}$$

if segment l is voiced **then**

Replace the cepstral coefficients $q \in \{q_{\mathrm{low,uv}}+1,\ldots,q_{\mathrm{low,v}}\}$

$$\bar{\phi}_{\mathrm{S},q} = \begin{cases} y_q & , q \in \{q_{\mathrm{low,uv}}+1,\ldots,q_{\mathrm{low,v}}\} \\ \phi_{\mathrm{S},q} & , \text{else.} \end{cases}$$

end if

Inverse cepstral transform (2.7)

$$\bar{\sigma}^2_{\mathrm{S},k} = \exp\left(\sum_{q=0}^{N-1} \bar{\phi}_{\mathrm{S},q}\, \mathrm{e}^{-\mathrm{j}2\pi kq/N}\right).$$

Estimate the clean speech spectral coefficients using the cepstrally modified spectral speech and noise powers $\bar{\sigma}^2_{\mathrm{S},k}$ and $\bar{\sigma}^2_{\mathrm{N},k}$.

end for

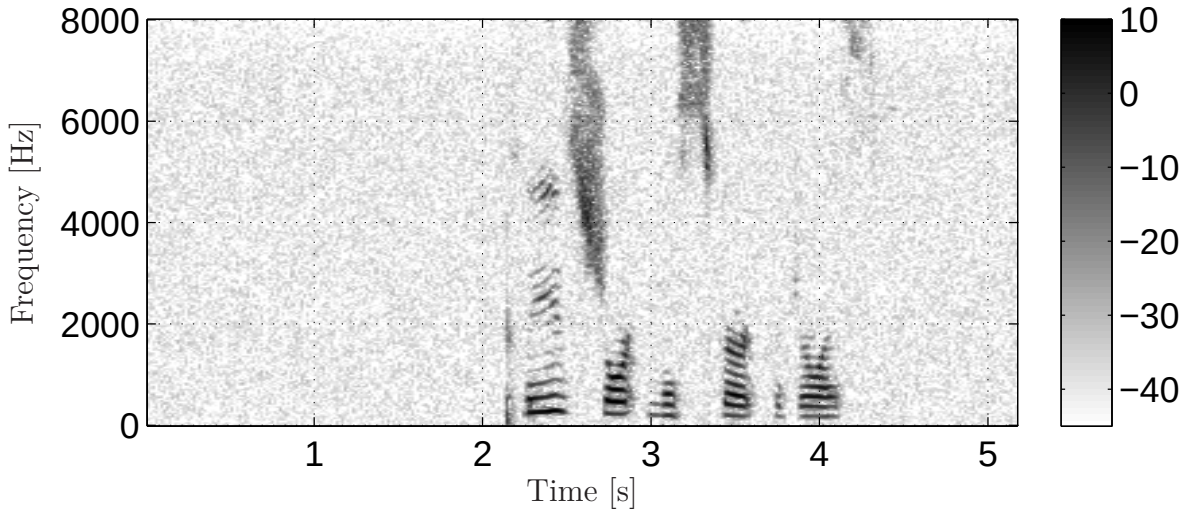


Figure 4.2: Spectrograms of enhanced speech using cepstral replacement. The clean speech of Figure 3.4 is disturbed by stationary white Gaussian noise at 0 dB segmental SNR. The noisy speech is given in Figure 3.5.

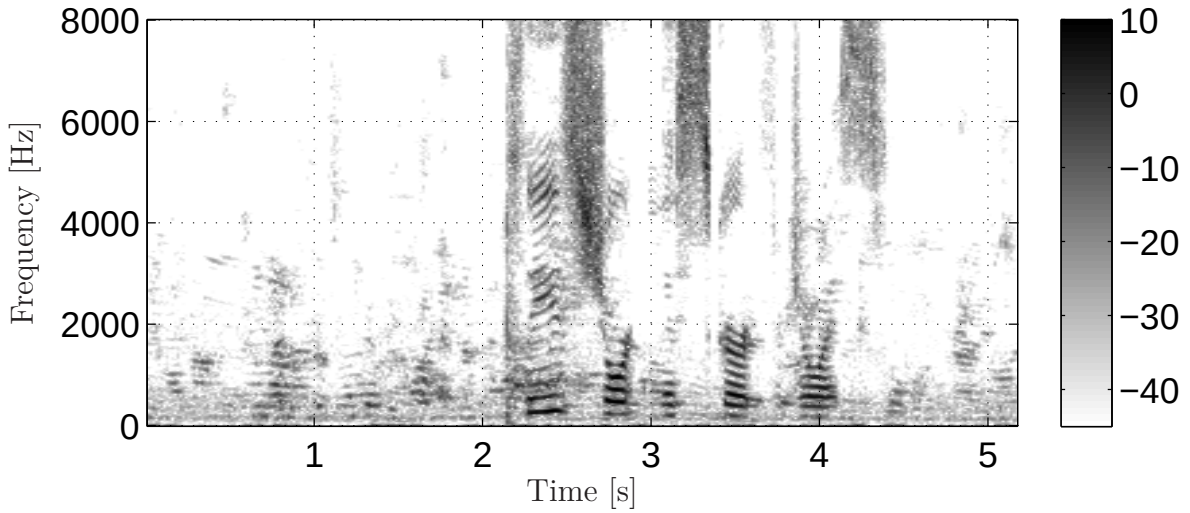


Figure 4.3: Spectrograms of enhanced speech using cepstral replacement. The clean speech of Figure 3.4 is disturbed by nonstationary babble noise at 0 dB segmental SNR.

4.4 Conclusions

In this chapter an alternative approach to a temporal smoothing of the cepstrum is presented. For single channel noise reduction the performance in terms of instrumental measures is better than a temporal smoothing of the cepstrum of spectral gain functions given in Section 3.1 and similar to Temporal Cepstrum Smoothing (TCS) for *a priori* Signal-to-Noise Ratio (SNR) estimation proposed in Section 3.2. The cepstral replacement approach is computationally much more expensive than the TCS approach. While for the TCS approach requires two Discrete Fourier Transforms (DFTs) for the cepstral transform and its inverse, the cepstral replacement approach requires one DFT for the cepstral transform of the noisy observation, and four DFTs for the cepstral transform of the spectral noise power, the spectral speech power and their inverses. However, informal listening has revealed a slight preference for the cepstral replacement approach for babble noise in low SNR scenarios. This can be attributed to the fact that the background noise is not temporally smeared when the instantaneous replacement approach is used.

Chapter 5

Speech Presence Probability Estimation

In this chapter we present an improved estimator for the speech presence probability at each time-frequency point in the short-time discrete Fourier transform domain. In contrast to existing approaches this estimator does not rely on an adaptively estimated and thus signal dependent *a priori* signal-to-noise ratio estimate. It therefore decouples the estimation of the speech presence probability from the estimation of the clean speech spectral coefficients in a speech enhancement task. Using both a fixed *a priori* signal-to-noise ratio and a fixed prior probability of speech presence, the proposed *a posteriori* speech presence probability estimator achieves probabilities close to zero for speech absence and probabilities close to one for speech presence. While state-of-the-art speech presence probability estimators use adaptive prior probabilities and signal-to-noise ratio estimates we argue that these quantities should reflect true *a priori* information that shall not depend on the observed signal. We present a detection theoretic framework for determining the fixed *a priori* signal-to-noise ratio. Also in this chapter, we derive the theoretical basis for *a posteriori* speech presence probability estimation based on a smoothed observation. The proposed estimator is conceptually simple and yields considerably less noise leakage and low speech distortions in both, stationary and nonstationary noise as compared to state-of-the-art estimators. Especially in babble noise, this results in large signal-to-noise ratio improvements. The results of this chapter are partly presented in [Gerkmann *et al*, 2008b] and [Gerkmann *et al*, 2010].

5.1 Introduction

For many short-time Discrete Fourier Transform (DFT) based speech processing systems an estimator for the Speech Presence Probability (SPP) in each time-frequency point is of great interest. For instance in speech enhancement clean-speech estimators,

such as the Wiener Filter (1.4), are often derived under the assumption that speech is actually present. Since this is neither true in speech pauses nor between the spectral harmonics of voiced speech sounds, the SPP should be taken into account [McAulay and Malpass, 1980, Ephraim and Malah, 1984, Malah *et al*, 1999, Cohen and Berdugo, 2001, Gerkmann *et al*, 2008b, Gerkmann *et al*, 2010]. SPP estimators are also of interest in multichannel speech enhancement to discard channels that are more severely disturbed than others [Gerkmann and Martin, 2006]. For clean-speech estimators, it is crucial that the SPP estimator does reliably recognize speech presence to avoid spectral distortions of low energy speech components. Most existing SPP estimators are designed in a way that they satisfy this demand, and yield high SPP estimates whenever speech is present. However, SPP estimators like [Ephraim and Malah, 1984, Malah *et al*, 1999], have the drawback that they usually do not yield small values for the SPP at time-frequency points where speech is absent, *e.g.* between the harmonics of voiced speech or even in speech pauses. The estimator in [Cohen and Berdugo, 2001] overcomes this problem by making the *a priori* SPP signal dependent. We argue that for *a posteriori* SPP estimation neither the *a priori* Signal-to-Noise Ratio (SNR) nor the *a priori* SPP should be adapted but represent true *a priori* knowledge and thus be independent of the observation. In this chapter we show that with fixed priors a better trade-off between speech distortions and noise reduction can be achieved as compared to competing state-of-the art estimators that adapt the priors, such as [Malah *et al*, 1999, Cohen and Berdugo, 2001].

The discussed SPP estimators require an estimate for the noise spectral power. However, state-of-the-art noise power spectral density estimators, like Martin's minimum statistics approach [Martin, 2001], are often based on the fact that noise is more stationary than speech. Consequently they are not capable of tracking instationarities such as high energy noise bursts of short duration that often occur in babble noise. This results in large SPP estimates during noise bursts, so that SPP estimators like [Ephraim and Malah, 1984, Malah *et al*, 1999, Cohen and Berdugo, 2001] exhibit a high false-alarm rate in nonstationary noise.

In this chapter we overcome this drawback by smoothing the *a posteriori* SNR in the cepstral domain. As a result, we present a new estimator for the *a posteriori* SPP which clearly outperforms state-of-the-art SPP estimators in nonstationary noise and also achieves better performance in stationary noise.

In the next section, we review the framework for *a posteriori* SPP estimation based on a smoothed *a posteriori* SNR, and show that for an adapted *a priori* SNR the *a posteriori* SNR yields only the prior SPP in speech absence. In Section 5.3 we argue that for *a posteriori* SPP estimation neither the *a priori* SNR nor the *a priori* SPP should be adapted, and present a framework to determine a fixed optimal *a priori* SNR that minimizes the false-alarm and missed-hit rates. While in Section 5.4 the *a posteriori* SNR is smoothed in the frequency domain, in Section 5.5 Temporal Cepstrum Smoothing

(TCS) of the *a posteriori* SNR is proposed. In Section 5.6 we show that *a posteriori* SPP based on TCS and fixed priors clearly outperforms the competing SPP estimators [Malah *et al*, 1999, Cohen and Berdugo, 2001, Gerkmann *et al*, 2008b] in nonstationary noise and also achieves better performance in stationary noise.

5.2 A posteriori SPP estimation

We now derive the *a posteriori* SPP given the *a posteriori* SNR. The *a posteriori* SNR is given as $\gamma_k = |S_k + N_k|^2/\sigma_{\mathrm{N},k}^2$ under speech presence $\mathcal{H}_{1,k}$, and $\gamma_k = |N_k|^2/\sigma_{\mathrm{N},k}^2$ in speech absence, denoted by $\mathcal{H}_{0,k}$. The noise power spectral estimate $\sigma_{\mathrm{N},k}^2 = \mathrm{E}\{|N_k|^2\}$ may be estimated using Martin's minimum statistics approach [Martin, 2001]. Using Bayes' theorem, the *a posteriori* probability of speech presence $P(\mathcal{H}_{1,k}|\,\gamma_k)$, can be obtained as

$$P(\mathcal{H}_{1,k}|\,\gamma_k) = \frac{P(\mathcal{H}_{1,k})\;\; p\Big(\gamma_k \;\Big|\; \mathcal{H}_{1,k}\Big)}{p(\gamma_k)} \tag{5.1}$$

$$= \frac{P(\mathcal{H}_{1,k})\;\; p\Big(\gamma_k \;\Big|\; \mathcal{H}_{1,k}\Big)}{P(\mathcal{H}_{1,k})\;\; p\Big(\gamma_k \;\Big|\; \mathcal{H}_{1,k}\Big) + P(\mathcal{H}_{0,k})\;\; p\Big(\gamma_k \;\Big|\; \mathcal{H}_{0,k}\Big)}. \tag{5.2}$$

Hence, the *a posteriori* SPP is fully defined by the likelihoods of speech presence $p\Big(\gamma_k \;\Big|\; \mathcal{H}_{1,k}\Big)$ and absence $p\Big(\gamma_k \;\Big|\; \mathcal{H}_{0,k}\Big)$ and the *a priori* SPP $P(\mathcal{H}_{1,k}) = 1 - P(\mathcal{H}_{0,k})$. The *a posteriori* SPP can be rewritten in terms of the generalized likelihood ratio as

$$P(\mathcal{H}_{1,k}|\,\gamma_k) = \frac{\Lambda_k}{1+\Lambda_k}, \tag{5.3}$$

where the generalized likelihood ratio Λ_k is defined as the weighted ratio of the likelihoods of speech presence and absence:

$$\Lambda_k = \frac{P(\mathcal{H}_{1,k}) \qquad p\Big(\gamma_k \;\Big|\; \mathcal{H}_{1,k}\Big)}{(1 - P(\mathcal{H}_{1,k}))\;\; p\Big(\gamma_k \;\Big|\; \mathcal{H}_{0,k}\Big)}. \tag{5.4}$$

The prior $P(\mathcal{H}_{1,k})$ can be used to bias the generalized likelihood ratio in favor of either speech presence ($P(\mathcal{H}_{1,k}) > 0.5$) or of speech absence ($P(\mathcal{H}_{1,k}) < 0.5$). All SPP estimators mentioned in Section 5.1 are implicitly or explicitly based on the generalized likelihood ratio.

5.2.1 Effects of a smoothed observation

The *a posteriori* SNR is defined as the periodogram of noisy speech $|Y_k|^2$ normalized on the noise power spectrum. Since the periodogram, as an estimate of the power spectrum, exhibits a large variance [Vary and Martin, 2006, Section 5.9], [Papoulis and Pillai, 2002, Section 12-2] the *a posteriori* SNR γ_k suffers from random fluctuations. When $P(\mathcal{H}_{1,k}|\,\gamma_k)$ is incorporated into a speech enhancement framework, random fluctuations in γ_k may result in spectral peaks in the enhanced signal that may be perceived as musical noise [Malah *et al*, 1999]. To reduce its variance, we propose to smooth the *a posteriori* SNR and denote the smoothed quantity as $\bar{\gamma}_k$.

As in [Ephraim and Malah, 1984, Malah *et al*, 1999, Cohen and Berdugo, 2001, Yu and Hansen, 2009, Gerkmann and Martin, 2006] we assume that Y_k is complex-Gaussian distributed, which results in a χ^2-distribution (2.11) with shape parameter $\mu = 1$ for the *a posteriori* SNR γ_k. For a smoothing in the frequency domain, it is well known, that the smoothed random variable remains approximately χ^2-distributed but with an increase in the degrees of freedom [Martin and Lotter, 2001, Martin, 2001]. The χ^2-distribution holds exactly if the averaged values of P are uncorrelated. For a cepstral smoothing the χ^2-distribution also holds, as shown in [Gerkmann and Martin, 2009] and Section 2.3.

As under speech absence we have $\mathrm{E}\{\gamma_k\} = 1$, we can write the likelihood of speech absence as

$$p\big(\bar{\gamma}_k \,\big|\, \mathcal{H}_{0,k}\big) = \frac{1}{\Gamma(\bar{\mu})}\bar{\mu}^{\bar{\mu}}\,\bar{\gamma}_k^{\bar{\mu}-1}\exp(-\bar{\mu}\,\bar{\gamma}_k)\ . \tag{5.5}$$

Assuming that speech and noise are uncorrelated, we have $\mathrm{E}\{\bar{\gamma}_k\} = 1 + \xi_k$ in speech presence, and thus

$$p\big(\bar{\gamma}_k \,\big|\, \mathcal{H}_{1,k}\big) = \frac{1}{\Gamma(\bar{\mu})}\left(\frac{\bar{\mu}}{1+\xi_k}\right)^{\bar{\mu}}\bar{\gamma}_k^{\bar{\mu}-1}\exp\!\left(-\bar{\mu}\,\frac{\bar{\gamma}_k}{1+\xi_k}\right)\ , \tag{5.6}$$

where $\xi_k = \sigma^2_{\mathrm{S},k}/\sigma^2_{\mathrm{N},k}$ is the *a priori* SNR and $\sigma^2_{\mathrm{S},k} = \mathrm{E}\{|S_k|^2\}$. The generalized likelihood ratio results in

$$\Lambda_k = \frac{P(\mathcal{H}_{1,k})}{1-P(\mathcal{H}_{1,k})}\cdot\left(\frac{1}{1+\xi_k}\right)^{\bar{\mu}}\exp\!\left(\frac{\xi_k}{1+\xi_k}\,\bar{\mu}\,\bar{\gamma}_k\right), \tag{5.7}$$

which is then used in (5.3) to compute the *a posteriori* SPP $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$.

The generalized likelihood ratio (5.4) is the ratio of the likelihoods (5.6) and (5.5) weighted by their priors. In order to illustrate the effect of the *a posteriori* SPP (5.3), Figure 5.1 shows the numerator and denominator of the generalized likelihood ratio

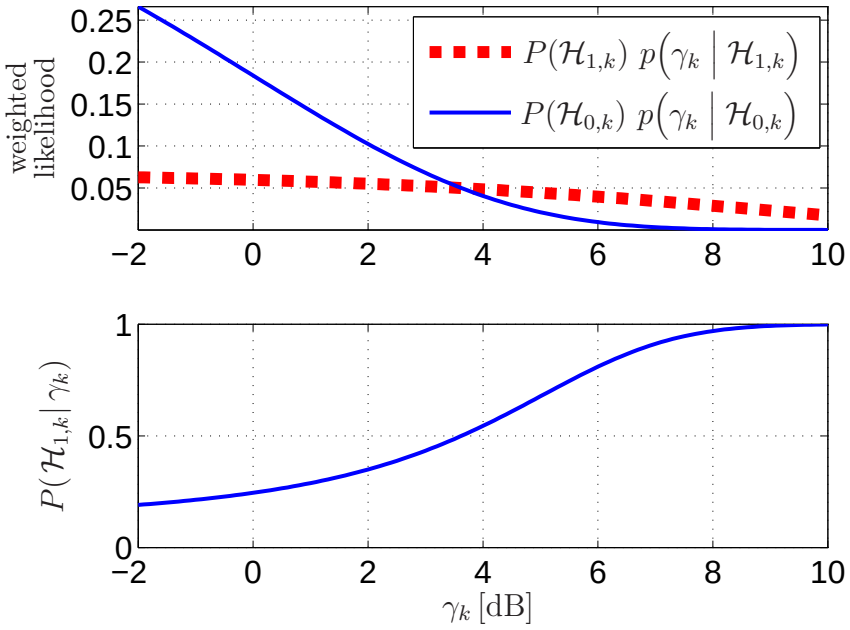


Figure 5.1: Numerator and denominator of the generalized likelihood ratio (5.4) and the resulting *a posteriori* SPP $P(\mathcal{H}_{1,k}|\,\gamma_k)$ for a nonsmoothed observation ($\mu = 1$), $10\log_{10}(\xi_k) = 8\,\text{dB}$, and $P(\mathcal{H}_{1,k}) = 0.5$. The prior $P(\mathcal{H}_{1,k})$ can be used to provide for an overall bias in favor of speech presence or absence. For $10\log_{10}(\gamma_k) > 10\log_{10}(\gamma_k^{\text{intersect}}) = 3.6\,\text{dB}$ the weighted ratio (5.4) of the likelihoods (5.6) and (5.5) is larger than one and the SPP, $P(\mathcal{H}_{1,k}|\,\gamma_k)$, is larger than 0.5.

(5.4) and the resulting SPPs for an *a priori* SNR of $10\log_{10}(\xi_k) = 8\,\text{dB}$ and $\mu = 1$ as a function of the *a posteriori* SNR. Note that while all computations are done in the linear domain, for the illustrations the *a posteriori* SNR is converted from linear scale to decibels, as $\gamma_k[\,\text{dB}] = 10\log_{10}(\gamma_k)$. The intersection of the weighted likelihoods $P(\mathcal{H}_{1,k})\ p(\gamma_k|\mathcal{H}_{1,k})$ and $(1 - P(\mathcal{H}_{1,k}))\,p(\gamma_k|\mathcal{H}_{0,k})$ occurs at

$$\gamma_k^{\text{intersect}} = \frac{1+\xi_k}{\xi_k}\log\left(\frac{1-P(\mathcal{H}_{1,k})}{P(\mathcal{H}_{1,k})}[1+\xi_k]\right) \tag{5.8}$$

and marks the point where the generalized likelihood ratio is $\Lambda_k = 1$ and where the resulting *a posteriori* SPP is $P(\mathcal{H}_{1,k}|\,\gamma_k) = 0.5$.

With [Gradshteyn and Ryzhik, 2000, (3.381.4)] the moments of a χ^2-distributed random variable can be computed, and it can be shown that the shape parameter after smoothing $\bar{\mu}$ is related to the ratio of the mean and variance after smoothing as

$$\bar{\mu} = (\mathrm{E}\{\bar{\gamma}_k\})^2 / \mathrm{var}\{\bar{\gamma}_k\} \; . \tag{5.9}$$

As argued in Section 2.3, from (5.9) it can be seen that the variance reduction of an unbiased smoothing with $\mathrm{E}\{\bar{\gamma}_k\} = \mathrm{E}\{\gamma_k\}$ necessarily results in an increase in the shape parameter of a χ^2-distributed random variable. Then, for a noise-only signal with $\mathrm{E}\{\bar{\gamma}_k\} = 1$ the shape parameter is simply given by the reciprocal of the reduced variance, as $\bar{\mu} = 1/\mathrm{var}\{\bar{\gamma}_k\}$. Thus, the shape parameter after smoothing can be obtained by measuring the variance of a noise only signal after smoothing as proposed in [Gerkmann *et al*, 2008b]. For a cepstral smoothing, $\bar{\mu}$ can be obtained as detailed in Algorithm 4.

For further theoretical analyses we employ the *false-alarm rate* and the *missed-hit rate*, as used in classical detection and estimation theory, *e.g.* [Van Trees, 1968, Ch. 2]. Interpreting the SPP estimator as a detector, we define the false-alarm rate as the probability that a noise-only bin yields an SPP higher than 0.5. Accordingly, the missed-hit rate is the probability that a bin that contains speech yields an SPP lower than 0.5.

The inherent variance reduction of the smoothing process results in less overlap of the likelihoods (5.5) and (5.6) and a steeper transition of the *a posteriori* SPP, as can be seen by comparing figures 5.1 and 5.2, as well as figures 5.4(a) and 5.4(b). The advantage of a steeper transition is that low values of $\bar{\gamma}_k$ yield a low SPP, when ξ_k is larger than a lower bound. The decreased overlap results in a lower false-alarm rate and a lower missed-hit rate, as shown next.

Using [Gradshteyn and Ryzhik, 2000, (3.381.3)], the false-alarm rate can be written as

$$P_{F,\bar{\mu}} = \int_{\gamma_k^{\mathrm{intersect}}}^{\infty} p(\bar{\gamma}_k|\mathcal{H}_{0,k}) \mathrm{d}\bar{\gamma}_k = \frac{\Gamma(\bar{\mu}, \bar{\mu}\,\gamma_k^{\mathrm{intersect}})}{\Gamma(\bar{\mu})} \, , \tag{5.10}$$

where $\gamma_k^{\mathrm{intersect}}$ is determined according to (5.8). For $\mu = 1$ this results in

$$P_{F,\mu=1} = \exp(-\gamma_k^{\mathrm{intersect}}) = \left(\frac{1 - P(\mathcal{H}_{1,k})}{P(\mathcal{H}_{1,k})} [1 + \xi_k] \right)^{-\frac{1+\xi_k}{\xi_k}} \; .$$

For $10\log_{10}(\xi_k) = 8\,\mathrm{dB}$ and $P(\mathcal{H}_{1,k}) = 0.5$ the false-alarm rate reduces from $P_{F,\mu=1} = 10\%$ for the unsmoothed case ($\mu = 1$) to $P_{F,\bar{\mu}} = 1\%$ for a smoothed observation with $\bar{\mu} = 5.1$. The missed-hit rate can be written as:

$$P_{M,\bar{\mu}} = \int_0^{\gamma_k^{\mathrm{intersect}}} p(\bar{\gamma}_k|\mathcal{H}_{1,k}) \mathrm{d}\bar{\gamma}_k = 1 - \frac{\Gamma(\bar{\mu}, \bar{\mu}\frac{\gamma_k^{\mathrm{intersect}}}{1+\xi_k})}{\Gamma(\bar{\mu})} \, . \tag{5.11}$$

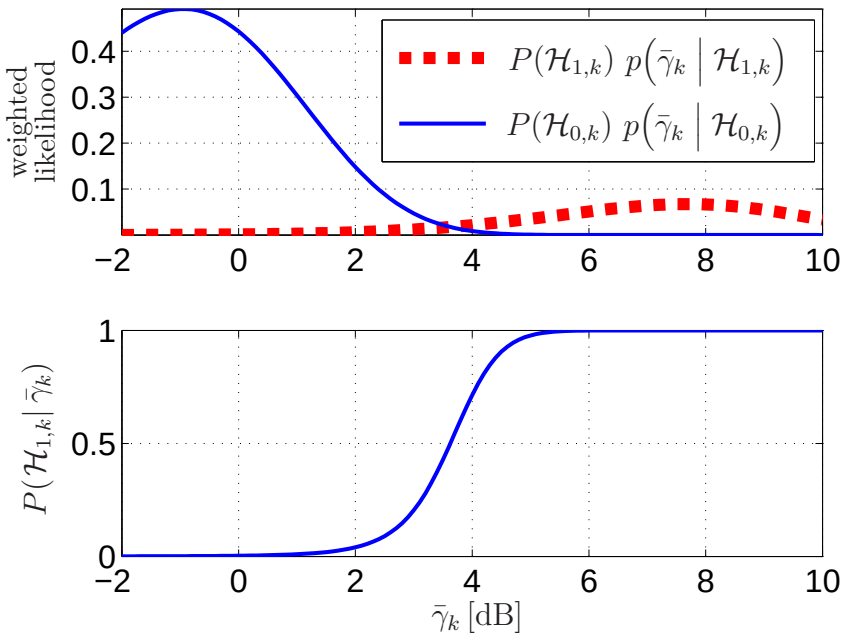


Figure 5.2: Numerator and denominator of the generalized likelihood ratio (5.4) and the resulting *a posteriori* SPP $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$ for a smoothed observation with $\bar{\mu} = 5.1$, $10\log_{10}(\xi_k) = 8\,\text{dB}$, and $P(\mathcal{H}_{1,k}) = 0.5$. Comparing this figure to Figure 5.1, it can be seen that smoothing results in less overlap of the likelihoods, which results in a lower false-alarm rate and missed-hit rate, as well as in a steeper transition of the *a posteriori* SPP $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$. The intersection of the likelihoods $10\log_{10}(\gamma_k^{\text{intersect}}) = 3.6\,\text{dB}$ is the same as in Figure 5.1.

For $\mu = 1$ this results in

$$P_{M,\mu=1} = 1 - \exp\left(-\frac{\gamma_k^{\text{intersect}}}{1+\xi_k}\right) = 1 - \left(\frac{1-P(\mathcal{H}_{1,k})}{P(\mathcal{H}_{1,k})}[1+\xi_k]\right)^{-\frac{1}{\xi_k}} .$$

For $10\log_{10}(\xi_k) = 8\,\text{dB}$ and $P(\mathcal{H}_{1,k}) = 0.5$ the missed-hit rate reduces from $P_{M,\mu=1} = 27\%$ considering a single bin ($\mu = 1$) to $P_{M,\bar{\mu}} = 2\%$ for a smoothed observation with $\bar{\mu} = 5.1$.

5.2.2 Drawbacks of an adapted *a priori* SNR

Besides the observation $\bar{\gamma}_k$ and the shape parameter $\bar{\mu}$, the estimate of $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$ depends on the *a priori* SPP $P(\mathcal{H}_{1,k})$ and the *a priori* SNR ξ_k ((5.3) and (5.7)). Since its introduction in [Ephraim and Malah, 1984], an estimate $\hat{\xi}_k$ of the *a priori* SNR is usually obtained using the decision-directed approach (3.14) [Ephraim and Malah, 1984, Malah *et al*, 1999, Cohen and Berdugo, 2001]. As the decision-directed approach (3.14) depends on an estimate of the clean speech $\widehat{S}_k$, the estimation of the SPP and the estimation of clean speech are coupled if the decision-directed approach is used for SPP estimation.

While the decision-directed approach and the *a priori* SNR estimator of Section 3.2 are powerful approaches to estimate the *a priori* SNR for filter gains, there is an intrinsic disadvantage to adapting the *a priori* SNR estimate for SPP estimation: at time-frequency points where speech is absent, the adapted *a priori* SNR is very small and thus the two likelihoods (5.5) and (5.6) that are compared in the generalized likelihood ratio (5.4) are approximately *the same* (cf. Figure 5.3). In this case the *a posteriori* SPP estimate $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$, does not make use of any information in the observation, but depends only on the *a priori* SPP $P(\mathcal{H}_{1,k})$. This can also be observed in Figure 5.4, where the *a posteriori* SPP is given for different *a priori* SNRs.

To overcome this problem Malah, Cox, and Accardi [Malah *et al*, 1999] suggested to perform two iterations on the SPP estimator: starting with a fixed $P(\mathcal{H}_{1,k}) = 0.5$, the resulting SPP estimate of the first iteration of (5.3) is used as a frequency dependent *a priori* SPP estimate $\widehat{P}(\mathcal{H}_{1,k}(l))$ in the second iteration. In figures 5.4(a) and 5.5 the *a posteriori* SPP $P(\mathcal{H}_{1,k}|\,\gamma_k)$, obtained with the conventional method and the iterative method proposed in [Malah *et al*, 1999] are given as a function of the *a posteriori* SNR. The second iteration of (5.3) causes a steeper transition of $P(\mathcal{H}_{1,k}|\,\gamma_k)$ from its minimum to its maximum, as may be seen by comparing figures 5.4(a) and 5.5. Note that in case that $\hat{\xi}_k$ is very small (*e.g.* $10\log_{10}(\xi_k) = -40\,\text{dB}$ in Figure 5.5), the resulting SPP still equals the prior $P(\mathcal{H}_{1,k})$, and the second iteration has no effect. This situation is somewhat improved, if we limit $\hat{\xi}_k$ to be larger than $10\log_{10}(\xi_{\min}) = -10\,\text{dB}$ as proposed

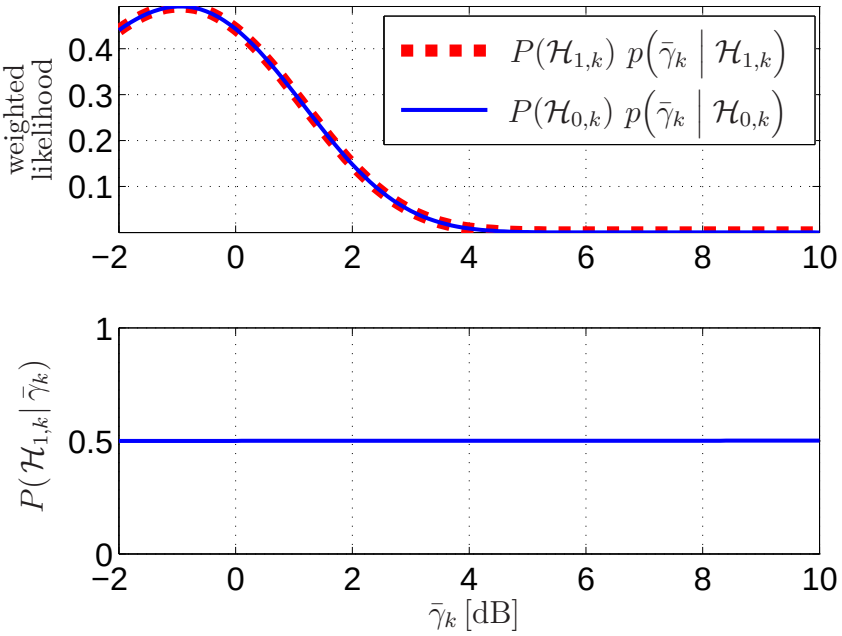
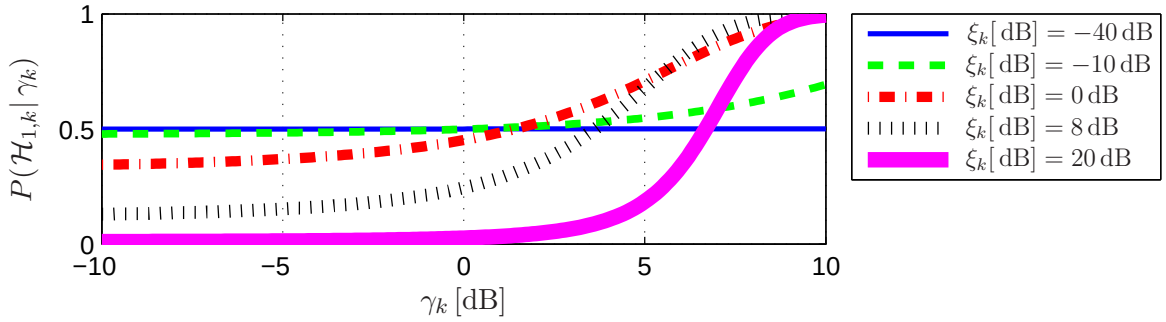
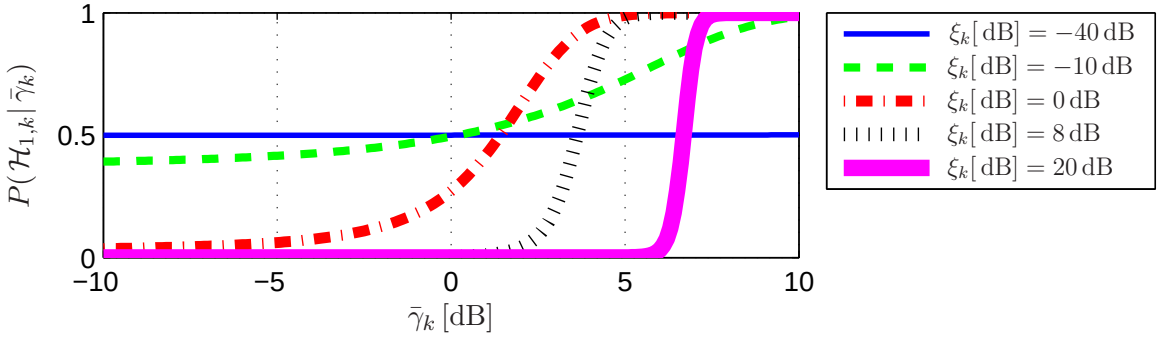


Figure 5.3: Numerator and denominator of the generalized likelihood ratio (5.4) and the resulting *a posteriori* SPP $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$ for a smoothed observation with $\bar{\mu} = 5.1$, $10\log_{10}(\xi_k) = -40\,\text{dB}$, and $P(\mathcal{H}_{1,k}) = 0.5$. For a small *a priori* SNR, *e.g.* $10\log_{10}(\xi_k) = -40\,\text{dB}$, the likelihoods overlap, and the *a posteriori* SPP $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$ yields only the *a priori* SPP $P(\mathcal{H}_{1,k}) = 0.5$ for all $\bar{\gamma}_k$.



(a) The conventional SPP estimator without smoothing ($\mu = 1$) and $P(\mathcal{H}_{1,k}) = 0.5$.



(b) The SPP estimator with a smoothed observation ($\bar{\mu} = 5.1$) and $P(\mathcal{H}_{1,k}) = 0.5$.

Figure 5.4: The speech presence probability $P(\mathcal{H}_{1,k}|\,\gamma_k)$, with and without smoothing. Smoothing the observation results in a steeper transition of the *a posteriori* SPP.

in [Malah *et al*, 1999]. The lower limit on $\hat{\xi}_k$ enables the SPP estimate to differ from $P(\mathcal{H}_{1,k})$ even in speech pauses, because the two likelihoods (5.6) and (5.5) cannot become identical. The second iteration emphasizes this difference, but the resulting SPP estimate is still far from zero for low SNR conditions, when $10\log_{10}\left(\hat{\xi}_k\right) = 10\log_{10}(\xi_{\min}) = -10\,\text{dB}$ (cf. Figure 5.5).

Cohen and Berdugo [Cohen and Berdugo, 2001] developed the idea of adapting the *a priori* SPP $P(\mathcal{H}_{1,k})$ further. Their approach exploits the correlation of speech presence in neighboring frequency bins of consecutive frames. This is done by taking local and global averages on the *a priori* SNR $\hat{\xi}_k$, as gained via the decision-directed approach (3.14). The averages are then mapped on values between 0 and 1 and reinterpreted as *a priori* SPPs. Since the resulting *a priori* SPP estimate $\widehat{P}(\mathcal{H}_{1,k})$ is mostly either very close to one or very close to zero, it dominates the *a posteriori* SPP. The likelihood-ratio in (5.4) has then only a minor effect.

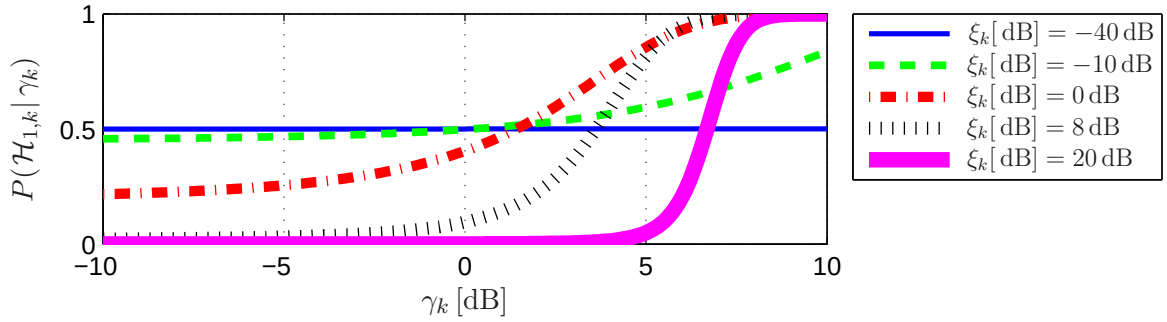


Figure 5.5: The speech presence probability $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$, without smoothing ($\mu = 1$) but with two iterations and an initial $P(\mathcal{H}_{1,k}) = 0.5$ as proposed in [Malah *et al*, 1999]. As compared to the case without smoothing and only one iteration given in Figure 5.4(a), the second iteration results in a steeper transition of the *a posteriori* SPP.

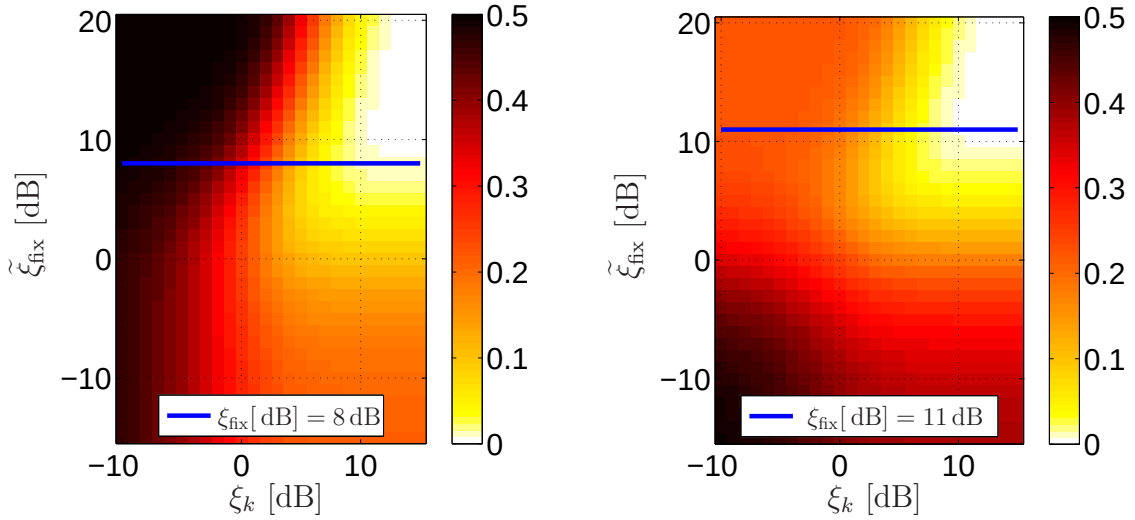
5.3 Fixed *a priori* SNR and *a priori* SPP

In this section, we argue that for *a posteriori* SPP estimation, the adaptation of the *a priori* SPP can be seen as only circumventing the true problem: the likelihoods $p\big(\bar{\gamma}_k \,\big|\, \mathcal{H}_{0,k}\big)$ and $p\big(\bar{\gamma}_k \,\big|\, \mathcal{H}_{1,k}\big)$ still tend to be equal in the absence of speech, which signifies a discrepancy in the basic probabilistic model. Instead, the *a priori* SPP and the *a priori* SNR should reflect true *a priori* knowledge and should not depend on the observation. Instead of adapting the *a priori* SPP and the *a priori* SNR, we therefore propose to use a fixed prior $P(\mathcal{H}_{1,k})$ and a constant ξ_{fix} that reflects the SNR that a typical speech sound would have *if speech were present* in the considered bin. This ξ_{fix} should be carefully chosen. If it is too high, the missed-hit rate increases, *i.e.* weak speech components are not recognized. If it is too low, the false-alarm rate increases, *i.e.* random fluctuations occur in $P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)$.

The optimal choice for ξ_{fix} is found by minimizing the average cost for a detection, which is denoted as the risk $\mathcal{R}$. With an assumed $\widetilde{\xi}_{\text{fix}}$, $\gamma_k^{\text{intersect}} = \frac{1+\widetilde{\xi}_{\text{fix}}}{\widetilde{\xi}_{\text{fix}}} \log\left(\frac{1-P(\mathcal{H}_{1,k})}{P(\mathcal{H}_{1,k})}[1+\widetilde{\xi}_{\text{fix}}]\right)$, (5.10), and (5.11), the risk combines the false-alarm rate $P_{F,\bar{\mu}}$ and the missed-hit rate $P_{M,\bar{\mu}}$, as

$$\mathcal{R}(\xi_k, \widetilde{\xi}_{\text{fix}}) = c_F\,[1 - P(\mathcal{H}_{1,k})]\,P_{F,\bar{\mu}}(\widetilde{\xi}_{\text{fix}}) + c_M\,P(\mathcal{H}_{1,k})\;P_{M,\bar{\mu}}(\widetilde{\xi}_{\text{fix}}, \xi_k)\,, \tag{5.12}$$

where c_F, c_M are the respective costs. The probabilities for correct detection are not considered in (5.12), since their cost is assumed to be zero. Note that the missed-hit



(a) $c_F = c_M$ (b) $c_F = 4\,c_M$

Figure 5.6: The risk $\mathcal{R}(\xi_k, \widetilde{\xi}_{\text{fix}})$ according to (5.12) as a function of the unknown *a priori* SNR ξ and the assumed $\widetilde{\xi}_{\text{fix}}$. A risk of zero corresponds to perfect detection. The larger the risk, the larger the probability of incorrectly assigning a bin to be speech or having missed a true speech bin. An integration of $\mathcal{R}(\xi_k, \widetilde{\xi}_{\text{fix}})$ along the horizontal line from $10\log_{10}(\xi_{\text{low}}) = -10\,\text{dB}$ to $10\log_{10}(\xi_{\text{up}}) = 15\,\text{dB}$ in the linear domain achieves the minimum overall risk. Here, $\bar{\mu} = 5.1$, and the *a priori* SPP is $P(\mathcal{H}_{1,k}) = 0.5$.

rate depends on the assumed *a priori* SNR $\widetilde{\xi}_{\text{fix}}$ and the unknown *a priori* SNR ξ_k. The false-alarm rate, however, is independent of the signal power and depends only on the assumed $\widetilde{\xi}_{\text{fix}}$. In Figure 5.6, the risk is illustrated for different costs c_F and c_M.

We find an optimal ξ_{fix} by minimizing the risk for all ξ between ξ_{low} and ξ_{up}.

$$\xi_{\text{fix}} = \arg\min_{\widetilde{\xi}_{\text{fix}}} \int_{\xi_{\text{low}}}^{\xi_{\text{up}}} \mathcal{R}(\xi_k, \widetilde{\xi}_{\text{fix}}) \mathrm{d}\xi_k \,. \tag{5.13}$$

We solve the integral numerically in the linear domain. When the resulting SPP estimate is applied to a speech enhancement framework, the costs for false-alarms c_F and missed-hits c_M control the trade-off between noise-leakage and speech distortions. These costs as well as the range of the integral in (5.13) can be adjusted, such that the performance of the SPP estimator is optimal for the application of interest. A

choice of $c_M = c_F = 1$ and zero cost for perfect detection minimizes the total probability of error [Van Trees, 1968]. As an example, for a smoothing that results in the shape parameter $\bar{\mu} = 5.1$, and the true SNR ranging from $10\log_{10}(\xi_{\text{low}}) = -10\,\text{dB}$ to $10\log_{10}(\xi_{\text{up}}) = 15\,\text{dB}$, the optimization (5.13) yields $10\log_{10}(\xi_{\text{fix}}) = 8\,\text{dB}$. Note that using a value of $10\log_{10}(\xi_{\text{fix}}) = 8\,\text{dB}$ in the generalized likelihood ratio does not mean that the *a posteriori* SNR has to be higher than $10\log_{10}(\bar{\gamma}_k) \approx 8\,\text{dB}$ to "detect" speech presence ($P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k) > 0.5$), but only higher than $10\log_{10}(\gamma_k^{\text{intersect}}) = 3.6\,\text{dB}$ (cf. figures 5.1, 5.4, and 5.4(b)).

As in [McAulay and Malpass, 1980] we assume that the probabilities of speech presence and speech absence in each bin are *a priori* equal, and thus set $P(\mathcal{H}_{1,k}) = 0.5$.

5.4 Smoothing the observation in the frequency domain

In this section we consider a frequency domain smoothing of the *a posteriori* SNR presented in [Gerkmann *et al*, 2008b]. To achieve the frequency domain smoothing, we calculate the smoothed observation over a time-frequency region in the neighborhood of the time-frequency point under consideration as

$$\bar{\gamma}_k(l) = \frac{1}{|\mathbb{K}| \cdot |\mathbb{L}|} \sum_{\substack{\kappa \in \mathbb{K} \\ \lambda \in \mathbb{L}}} \gamma_\kappa(\lambda)\,. \tag{5.14}$$

Here, $\mathbb{K}$ is the set of adjacent frequency bins, $\mathbb{L}$ is the set of successive time frames, and $|\mathbb{K}| \cdot |\mathbb{L}|$ is the number of spectral bins used for averaging.

The shape parameter $\bar{\mu}$ is increased as compared to the unsmoothed case where only one spectral bin is considered and $\mu = 1$. If the time-frequency points $\mathbb{K}\times\mathbb{L}$ are uncorrelated, the shape parameter is $\bar{\mu} = |\mathbb{K}| \cdot |\mathbb{L}|$. However, due to the overlapping tapered analysis windows w_n in (1.1), adjacent time-frequency points are usually correlated. Then, the shape parameter after smoothing can be determined empirically by smoothing a noise only signal and measuring its variance. As for a noise only signal $\mathrm{E}\{\bar{\gamma}_k\} = 1$, the shape parameter after smoothing can be obtained using (5.9).

Without loss of generality, we discuss the smoothing defined in (5.14) in the context of a causal system and choose $\mathbb{K}$ and $\mathbb{L}$ according to Figure 5.7. The parameters Δl and Δk defined in Figure 5.7 should be chosen large enough to ensure a low false-alarm rate, but small enough to preserve the fine structure of speech. In [Cohen and Berdugo, 2001] and [Sørensen and Andersen, 2005] the combination of two initial SPPs $P(\mathcal{H}_{1,k}|\bar{\gamma}_{\text{local},k})$ and $P(\mathcal{H}_{1,k}|\bar{\gamma}_{\text{global},k})$ has been successfully applied. These initial SPPs are based on different averaging windows. $P(\mathcal{H}_{1,k}|\bar{\gamma}_{\text{global},k})$ is based on a relatively large averaging

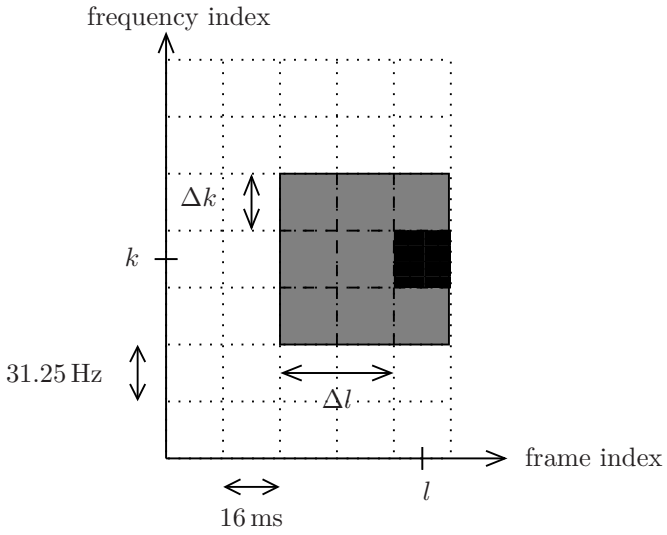


Figure 5.7: Illustration of the computation of the smoothed observation $\bar{\gamma}_k(l)$ via (5.14) in the time-frequency domain for $f_\mathrm{s} = 16\,\mathrm{kHz}$, $N = 512$ and a segment overlap of $L/N = 50\%$ in (1.1). The current bin k, l is marked black. The gray area illustrates the neighboring bins used for the smoothing, giving $|\mathbb{K}| \cdot |\mathbb{L}| = [\Delta l + 1] \cdot [2\Delta k + 1]$ bins.

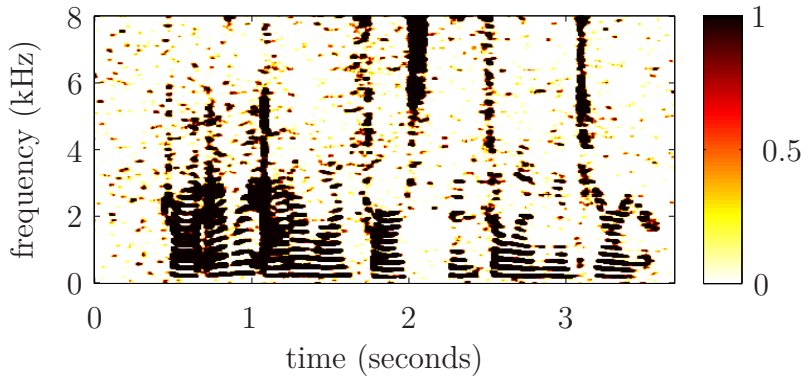
window. Thus, its variance is greatly reduced, but the fine structure of the speech signal is lost (see example in Figure 5.8(b)). On the other hand $P(\mathcal{H}_{1,k}|\bar{\gamma}_{\mathrm{local},k})$ is based on a much smaller averaging window. It has a high variance but is able to resolve the fine structure of the speech signal (see example in Figure 5.8(a)). As in [Cohen and Berdugo, 2001, Sørensen and Andersen, 2005] we propose to combine two initial SPPs such that the final SPP estimator yields values close to one only if the global *and* the local SPP have values close to one. This is achieved via a multiplicative combination [Cohen and Berdugo, 2001, Sørensen and Andersen, 2005], as:

$$\mathcal{P}_k = P(\mathcal{H}_{1,k}|\bar{\gamma}_{\mathrm{local},k}) \cdot P(\mathcal{H}_{1,k}|\bar{\gamma}_{\mathrm{global},k})\,, \tag{5.15}$$

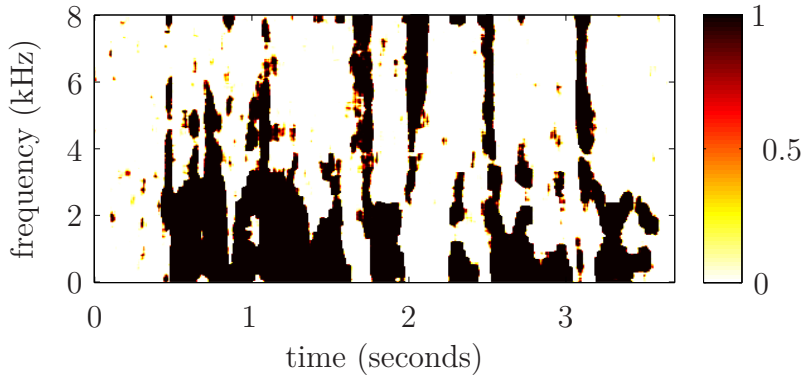
where $\mathcal{P}_k$ is an estimate of the *a posteriori* SPP. In Figure 5.8(c) it can be seen that the combined SPP (5.15) based on the local and global averages presented in Figure 5.8(a) and Figure 5.8(b), has a low variance but resolves the fine structure of speech.

For our purposes, the following averaging parameters were found to yield a good trade-off between tempo-spectral resolution, missed-hit rate, and false-alarm rate of the proposed SPP estimator. For the temporal smoothing, we propose to average over $\bar{T} = 64\,\mathrm{ms}$ of speech. With

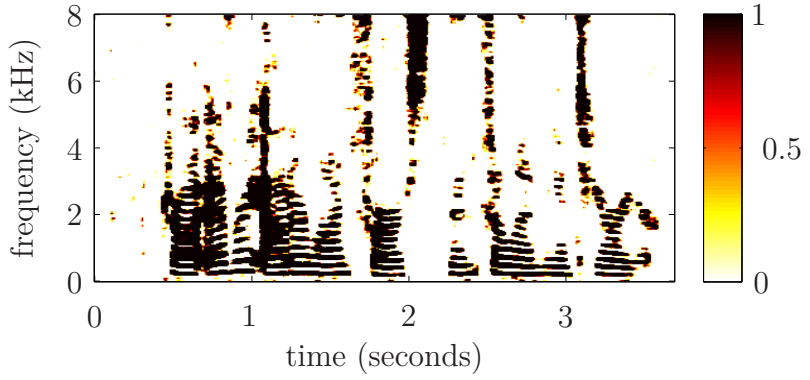
$$\Delta l = (\bar{T} - T_\mathrm{seg})/T_\mathrm{shift}\,, \tag{5.16}$$



(a) $P(\mathcal{H}_{1,k}|\bar{\gamma}_{\mathrm{local},k})$



(b) $P(\mathcal{H}_{1,k}|\bar{\gamma}_{\mathrm{global},k})$



(c) $\mathcal{P}_k = P(\mathcal{H}_{1,k}|\bar{\gamma}_{\mathrm{local},k}) \cdot P(\mathcal{H}_{1,k}|\bar{\gamma}_{\mathrm{global},k})$

Figure 5.8: The local SPP (a), the global SPP (b), and their product (c) for an exemplary speech signal disturbed by pink noise at 0 dB segmental SNR.

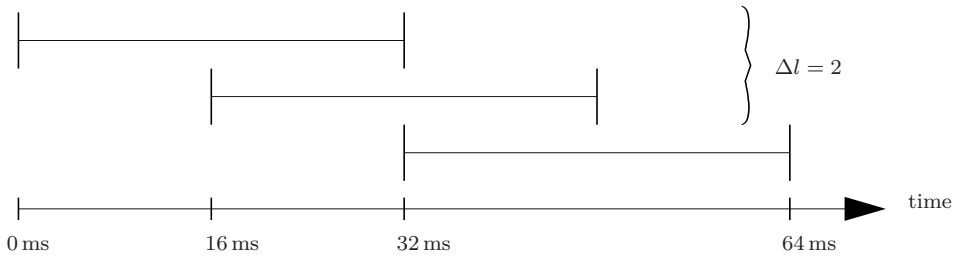


Figure 5.9: Illustration of overlapping time segments. With an analysis window of length 32 ms and a frame-shift of 8 ms the overall time averaging window of 64 ms requires $\Delta l + 1 = 3$ segments.

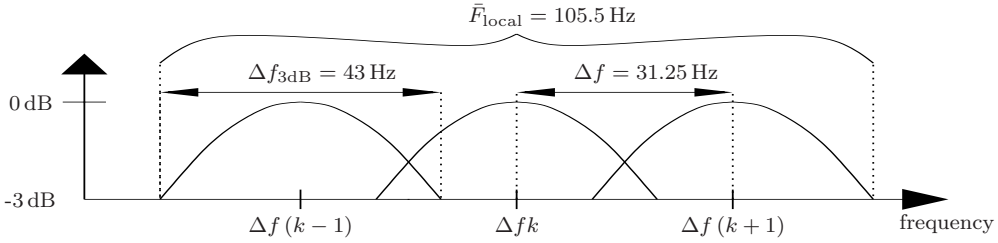


Figure 5.10: Illustration of frequency averaging. With a distance between frequency bands of 31.25 Hz and a 3 dB mainlobe bandwidth of 43 Hz the overall frequency averaging window of 105.5 Hz requires $2\,\Delta k + 1 = 3$ bins.

the analysis segment length of $T_{\text{seg}} = N/f_{\text{s}} = 32\,\text{ms}$ and a segment shift of $T_{\text{shift}} = L/f_{\text{s}} = 16\,\text{ms}$ (50% overlap), this results in $\Delta l = 2$ (cf. Figure 5.7 and Figure 5.9). For the smoothing along frequency we have

$$\Delta k_{\Xi} = \frac{1}{2}(\bar{F}_{\Xi} - \Delta f_{3\text{dB}})/\Delta f\,, \tag{5.17}$$

with a frequency bin distance of $\Delta f = 1/T_{\text{seg}} = 31.25$ Hz, and a 3 dB mainlobe bandwidth of the Hann window of approximately $\Delta f_{3\text{dB}} \approx 43\,\text{Hz}$. In (5.17) and below, Ξ stands for either the local or the global average. For the local average we want to apply only little smoothing to preserve the fine structure of speech. We propose to average over a frequency window of $\bar{F}_{\text{local}} = 105.5\,\text{Hz}$ which results in $\Delta k_{\text{local}} = 1$ (cf. Figure 5.7 and Figure 5.10). For the global average we want to have a relatively large frequency window to reduce fluctuations in the observation. We choose a frequency window of 543 Hz that results in $\Delta k_{\text{global}} = 8$.

Using these values for the averaging in the time-frequency plane and using $c_F = c_M = 1$, $10\log_{10}(\xi_{\text{low}}) = -10\,\text{dB}$, and $10\log_{10}(\xi_{\text{up}}) = 15\,\text{dB}$ for the computation of the optimal ξ_{fix}, we get the parameters given in Table 5.1. The parameters are determined as summarized in Algorithm 6. Note that the resulting parameters $\bar{\mu}_{\Xi}$ and $\xi_{\text{fix},\Xi}$ are insensitive

window-overlap	Ξ	Δk_Ξ	Δl	$\lvert\mathbb{K}\rvert \cdot \lvert\mathbb{L}\rvert$	$\bar{\mu}_\Xi$	$\xi_{\text{fix},\Xi}[\,\text{dB}]$
50%	local	1	2	9	5.4	8 dB
	global	8	2	51	25.7	3 dB
75%	local	1	4	15	5.1	8 dB
	global	8	4	85	24.7	3 dB

Table 5.1: The parameter settings for the proposed SPP estimator with $\bar{T} = 64$ ms, $\bar{F}_{\text{local}} = 105.5$ Hz and $\bar{F}_{\text{global}} = 543$ Hz. Ξ stands for either the local or the global average. The parameters are determined for a Hann window with 50% overlap and 75% overlap, respectively. The fixed SNR is optimized for a range of $10\log_{10}(\xi_{\text{low}}) = -10\,\text{dB}$ to $10\log_{10}(\xi_{\text{up}}) = 15\,\text{dB}$.

to different choices of window overlaps, if the number of time frames Δl is chosen accordingly, as the difference in correlation is considered in $\bar{\mu}$ (cf. Table 5.1). The algorithm for estimating the SPP is summarized in Algorithm 7.

Algorithm 6 Determination of the parameters in Table 5.1. Ξ stands for either the local or the global average.

1: choose averaging window, *e.g.* $\bar{T} = 64\,\text{ms}$, $\bar{F}_{\text{global}} = 543$ Hz, and $\bar{F}_{\text{local}} = 105.5$ Hz
2: compute Δl and Δk_Ξ via (5.16), (5.17)
3: compute the respective number of bins, $\lvert\mathbb{K}\rvert \cdot \lvert\mathbb{L}\rvert = (\Delta l + 1) \cdot (2\Delta k_\Xi + 1)$
4: empirically determine the degrees of freedom by smoothing a noise only signal and using (5.9), $\bar{\mu}_\Xi = (\text{E}\{\bar{\gamma}_{\Xi,k}\})^2/\text{var}\{\bar{\gamma}_{\Xi,k}\}$
5: compute the optimal *a priori* SNR, $\xi_{\text{fix},\Xi}$ (5.13)

Algorithm 7 Proposed SPP estimation algorithm based on frequency domain smoothing. Ξ stands for either the local or the global average.

1: **for all** signal segments l **do**
2: compute smoothed observation $\bar{\gamma}_{k,\Xi}$ (5.14)
3: compute $P(\mathcal{H}_{1,k}|\bar{\gamma}_{\Xi,k})$ via (5.3) and (5.7) using $\xi_{\text{fix},\Xi}$ and $P(\mathcal{H}_{1,k}) = 0.5$
4: compute the overall SPP: $\mathcal{P}_k = P(\mathcal{H}_{1,k}|\bar{\gamma}_{\text{global},k}) \cdot P(\mathcal{H}_{1,k}|\bar{\gamma}_{\text{local},k})$
5: **end for**

5.5 Smoothing the observation in the cepstral domain

Smoothing in the time-frequency domain is always a trade-off between tempo-spectral resolution and a reduction of outliers. A reduction of the temporal resolution may smear speech onsets, while a reduction of the spectral resolution may dissallow resolving spectral harmonics of voiced speech sounds. To obtain *a posteriori* SPP estimates that exhibit a low variance and still resolve the spectral harmonics, for the frequency domain smoothing of [Cohen and Berdugo, 2001, Sørensen and Andersen, 2005, Gerkmann *et al*, 2008b] and Section 5.4, a multiplicative combination of initial *a posteriori* SPP estimates based on local and global averages are required.

In the cepstral domain, we can exploit *a priori* knowledge about which cepstral coefficients are likely to represent speech and apply a selective smoothing that results in an effective reduction of spectral outliers while preserving the speech spectral envelope. Thus, we propose to smooth the *a posteriori* SPP estimate in the cepstral domain via Algorithm 1. The bias correction factor $\mathcal{B}$ in Algorithm 1 is computed using Algorithm 4 where we assume the *a posteriori* SNR before smoothing γ_k to be χ^2-distributed with $\mu = 1$. Algorithm 1 is also used to estimate the shape parameter after smoothing $\bar{\mu}$ that is needed in (5.7).

The adaptive smoothing factor α_q in Algorithm 1 is obtained as in (3.4). For the recursive smoothing constant α_q^{const} in (3.4), we choose:

$$\alpha_q^{\text{const}} = \begin{cases} 0.2 & , q \in \{0, ..., 2\} \\ 0.4 & , q \in \{3, ..., 23\} \\ 0.997 & , q \in \{24, ..., 256\}\,, \end{cases} \tag{5.18}$$

where $\mathbb{Q}_{\text{pitch}}$ is estimated using (3.2) and (3.3). With the given α_q^{const} and assuming that the true, unknown SNR ranges from -10 dB to 25 dB, using (5.13) the optimal fixed *a priori* SNR is found to be $10\log_{10}(\xi_{\text{fix}}) = 9\,\text{dB}$. The used parameters are summarized in Table 5.2.

Smoothing factor for $\mathbb{Q}_{\text{pitch}}$ (3.4)	α_{pitch}	$= 0.5$
Threshold for voiced/unvoiced decision (3.3)	Λ^{thr}	$= 0.2$
Lower bound for the q_0 search (3.2)	q_{low}	$= \frac{f_{\text{s}}}{300\,\text{Hz}}$
Upper bound for the q_0 search (3.2)	q_{high}	$= \frac{f_{\text{s}}}{70\,\text{Hz}}$
Margin for $\mathbb{Q}_{\text{pitch}}$ (3.3)	Δq_0	$= 2$
Length of the cepstral low-pass (3.1)	τ_{H}	$= \frac{f_{\text{s}}}{2000\,\text{Hz}}$
Smoothing constant for (3.4)	β	$= 0.96$
Sampling rate	f_{s}	$= 16\,\text{kHz}$
Lower limit for the integral in (5.13)	$10\log_{10}(\xi_{\text{low}})$	$= \text{-}10\,\text{dB}$
Upper limit for the integral in (5.13)	$10\log_{10}(\xi_{\text{up}})$	$= 25\,\text{dB}$
Optimal fixed *a priori* SNR	$10\log_{10}(\xi_{\text{fix}})$	$= 9\,\text{dB}$
A priori SPP	$P(\mathcal{H}_{1,k})$	$= 0.5$

Table 5.2: Parameters for a TCS of the *a posteriori* SNR.

As in the cepstral domain a selective smoothing is applied where only little smoothing is applied to the speech related cepstral coefficients and a strong smoothing to the remaining cepstral coefficients, the multiplicative combination (5.15) of estimators based on local and global averages as applied in algorithms based on frequency domain smoothing [Cohen and Berdugo, 2001, Sørensen and Andersen, 2005] and Section 5.4 is not necessary. Thus, after smoothing the *a posteriori* SNR in the cepstral domain we obtain the *a posteriori* SPP

$$\mathcal{P}_k = P(\mathcal{H}_{1,k}|\,\bar{\gamma}_k)\,. \tag{5.19}$$

In Section 5.6 we show that a temporal cepstrum smoothing clearly outperforms frequency domain smoothing in nonstationary noise and also achieves better performance in stationary noise.

5.6 Experimental results

In this section, we compare different state-of-the-art *a posteriori* SPP estimators to the proposed estimators with fixed priors using frequency domain smoothing and cepstrum domain smoothing.

In figures 5.11 and 5.12 the spectrograms and resulting SPP estimates for stationary white noise and a female speaker are shown, while in figures 5.13 and 5.14 the results for nonstationary babble noise and a male speaker are given. It can be seen that the estimator proposed in [Malah *et al*, 1999] does not yield SPP estimates close to zero in speech absence. Furthermore, dark speckles in the gray spectral regions where no speech is present indicate a large amount of spectral outliers that may yield musical noise in a speech enhancement task. This undesired behavior is overcome by the estimator of [Cohen and Berdugo, 2001] and the proposed approaches. However, it can be seen that for white noise (Figure 5.12) the estimator of Cohen *et al.* does not yield high SPPs for the fricatives at $l = 63$ and $l = 192$ which results in larger speech distortions as compared to the proposed estimators. For the male speaker in Figure 5.14 it can be seen that the approach of Cohen *et al.* and the proposed frequency domain smoothing of Section 5.4 do not resolve the spectral harmonics, which results in large false alarm rates. The proposed cepstral domain smoothing approach of Section 5.5 does not only yield low SPP estimates in speech absence, but also resolves the spectral harmonics and yields larger SPPs for fricatives in white noise as compared to the approach of [Cohen and Berdugo, 2001]. Thus, it is expected to yield less speech distortions and less noise leakage, which will be confirmed by instrumental measures next.

Inspired by [Hu and Wang, 2004] we evaluate the SPP estimators in terms of Speech Distortions (SD) and Noise Leakage (NL), which can be seen as measures for missed-hit rate and false-alarm rate, respectively. As in [Erkelens *et al*, 2007b] we create an ideal binary speech presence mask $\mathcal{P}_{\mathrm{id},k}$ from the clean speech signal S_k that contains ones at all short-time DFT bins where the energy is no less than 50 dB below the maximum bin energy in the particular speech signal. We then compute two error-signals, $E_{\mathrm{SD},k}$ and $E_{\mathrm{NL},k}$ as:

$$E_{\mathrm{SD},k} = \max\left\{\mathcal{P}_{\mathrm{id},k} - \mathcal{P}_k,\, 0\right\} \cdot S_k\,, \tag{5.20}$$

$$E_{\mathrm{NL},k} = \max\left\{\mathcal{P}_k - \mathcal{P}_{\mathrm{id},k},\, 0\right\} \cdot N_k\,, \tag{5.21}$$

where $\mathcal{P}_k$ is the estimated *a posteriori* SPP obtained *e.g.* as (5.15) or (5.19). $E_{\mathrm{SD},k}$ contains those speech bins that are marked as speech by the ideal mask $\mathcal{P}_{\mathrm{id},k}$, but are attenuated by the SPP estimator $\mathcal{P}_k$. $E_{\mathrm{NL},k}$ contains those noise bins that are marked as noise by the ideal mask $\mathcal{P}_{\mathrm{id},k}$, but are not fully suppressed by the SPP estimator $\mathcal{P}_k$. These error signals are then related to the ideal speech signal $S_{\mathrm{id},k}$ and the corresponding ideal noise signal $N_{\mathrm{id},k}$, which are gained as:

$$S_{\mathrm{id},k} = \mathcal{P}_{\mathrm{id},k} S_k\,, \tag{5.22}$$

$$N_{\mathrm{id},k} = (1 - \mathcal{P}_{\mathrm{id},k}) N_k\,. \tag{5.23}$$

After taking the inverse Fourier transform and reconstructing the time signal by overlapping and adding the signal segments we get the time domain signals $e_{\mathrm{SD}}(\tau)$, $e_{\mathrm{NL}}(\tau)$,

$s_{\mathrm{id}}(\tau)$ and $n_{\mathrm{id}}(\tau)$. The final measures for speech distortions and noise leakage are then gained as:

$$\mathrm{SD} = \frac{\sum_\tau e_{\mathrm{SD}}^2(\tau)}{\sum_\tau s_{\mathrm{id}}^2(\tau)}, \tag{5.24}$$

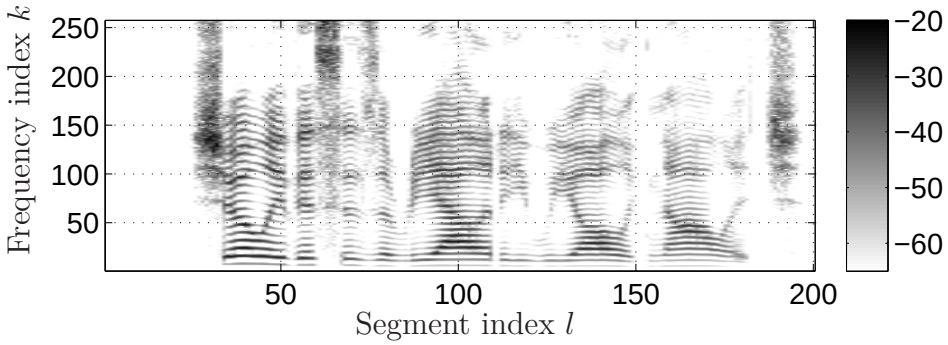
$$\mathrm{NL} = \frac{\sum_\tau e_{\mathrm{NL}}^2(\tau)}{\sum_\tau n_{\mathrm{id}}^2(\tau)}. \tag{5.25}$$

The measure for speech distortions SD indicates the percentage of the speech energy that the corresponding SPP estimator neglects while the measure for noise leakage NL indicates how much energy from the noise-only bins is not attenuated (in percent). Thus, SD equals 100% if all speech coefficients indicated by the ideal mask $\mathcal{P}_{\mathrm{id}}$ are attenuated by $\mathcal{P}_k$ and SD = 0% if $\mathcal{P}_k = 1$ wherever $\mathcal{P}_{\mathrm{id},k} = 1$. The NL equals 0% if $\mathcal{P}_k = 0$ for all noise-only bins.

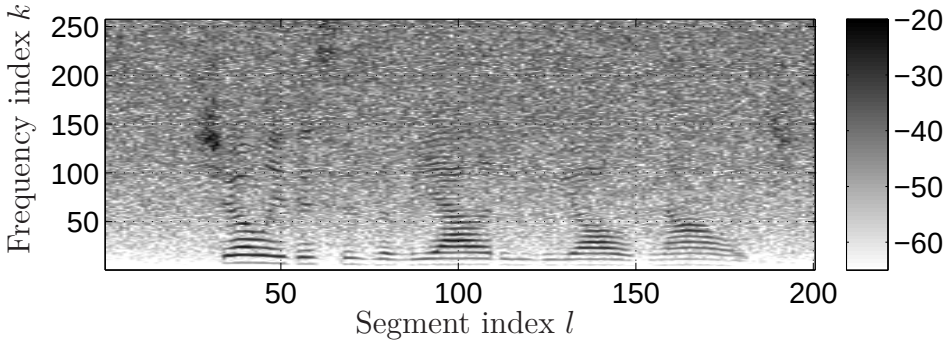
Furthermore, we quantify the segmental SNR improvement when the SPP estimate $\mathcal{P}_k$ is applied multiplicatively to noisy speech coefficients Y_k as

$$\widehat{S}_k = \mathcal{P}_k Y_k \,. \tag{5.26}$$

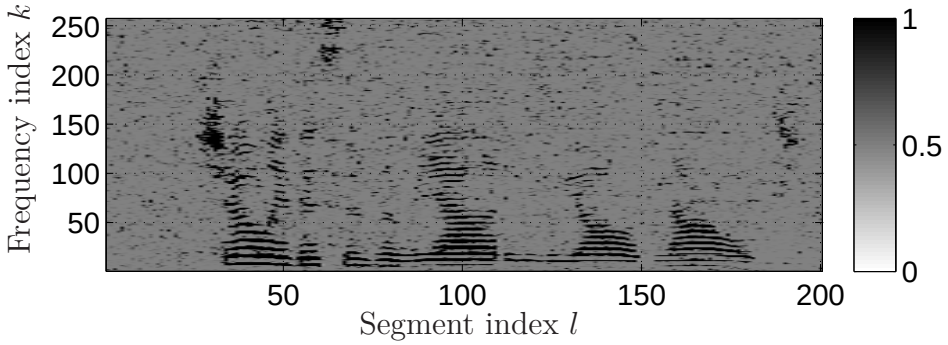
We process 320 speech samples from dialect region 6 of the TIMIT database [Garofolo, 1988] which are phonetically balanced and are from both male and female speakers. The speech is disturbed by white Gaussian noise, babble noise inside a crowded restaurant, and nonstationary traffic noise at a busy street, respectively. For the short-time Fourier analysis (1.1) we use Hann windows w_n with a length of 32 ms and 50% overlap. The spectral noise power is estimated using the minimum statistics approach [Martin, 2001]. The experimental results for input segmental SNRs between -10 and 15 dB are given in Figure 5.15. It can be seen that the proposed approaches that use both a fixed *a priori* SPP and a fixed *a priori* SNR yield less noise leakage than the competing approaches [Malah *et al*, 1999, Cohen and Berdugo, 2001] for all considered input SNRs and noise types. In terms of speech distortions, the proposed approaches perform similar to the approach of [Cohen and Berdugo, 2001] for babble noise and traffic noise, and slightly better for white noise. The estimator of [Malah *et al*, 1999] exhibits even lower speech distortions as it does not yield values close to zero in speech absence. Consequently, it gives a large noise leakage and results in a poor SNR improvement. While the proposed frequency domain smoothing of Section 5.4 yields similar results in terms of the SNR improvement as compared to [Cohen and Berdugo, 2001], the proposed TCS based estimator outperforms the competing estimators especially in babble noise. At 0 dB SNR, the segmental SNR indicates a considerable gain of approximately 1.5 dB as compared the competing algorithms in babble noise and a gain of approximately 0.5 dB in stationary white Gaussian noise.



(a) Clean signal.

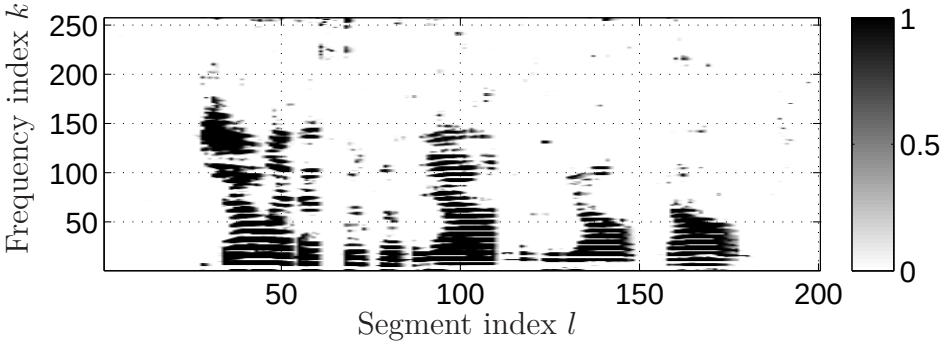


(b) Noisy signal.

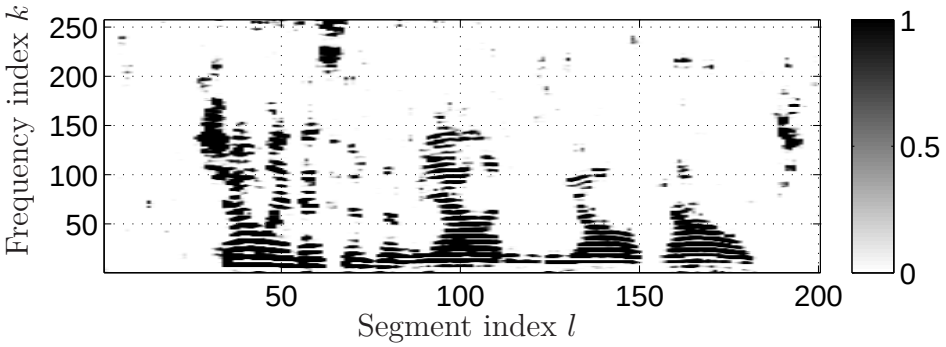


(c) *A posteriori* SPP according to [Malah *et al*, 1999].

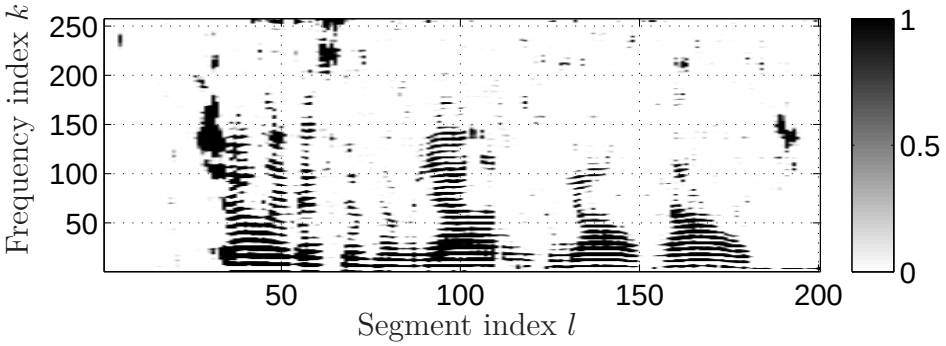
Figure 5.11: Clean speech (a), noisy speech (b), and the resulting *a posteriori* SPP estimate using [Malah *et al*, 1999] (c) for the sentence "Surely this is a reality we all acknowledge" spoken by a female speaker disturbed by additive white noise at 0 dB input segmental SNR. The signals in the spectrograms (a) and (b) have been pre-emphasized for a better visualization of high-frequency components. The estimator [Malah *et al*, 1999] in (c) does not yield SPP estimates close to zero in speech absence.



(a) *A posteriori* SPP according to [Cohen and Berdugo, 2001].

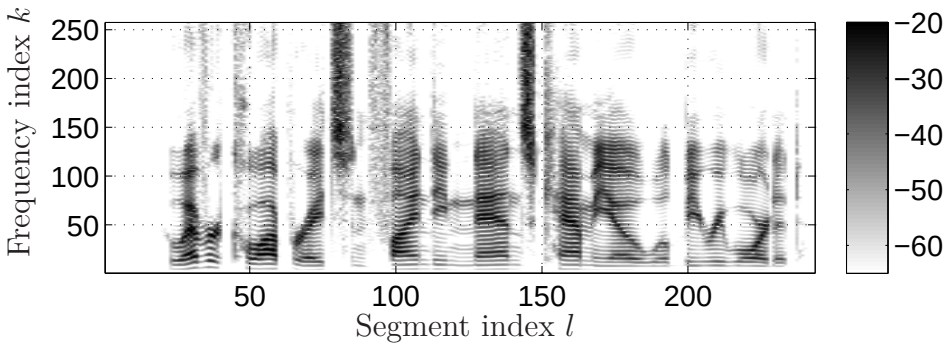


(b) Frequency domain smoothing proposed in Section 5.4 [Gerkmann *et al*, 2008b].

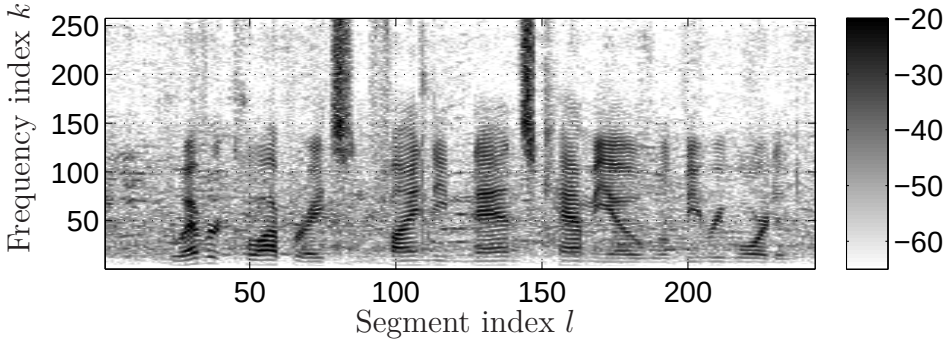


(c) Cepstral domain smoothing proposed in Section 5.5 [Gerkmann *et al*, 2010].

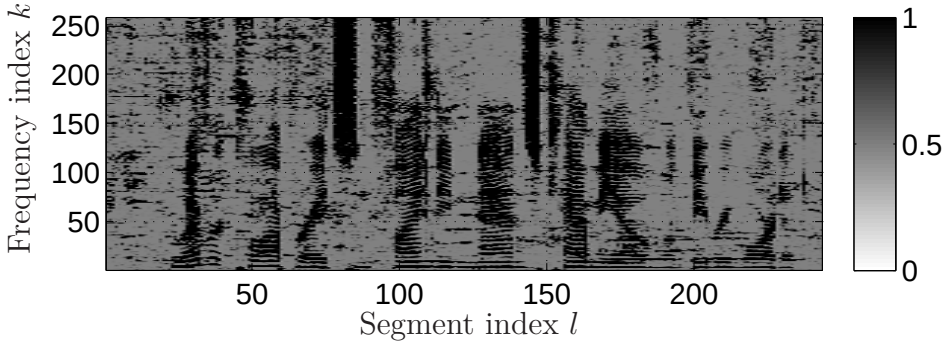
Figure 5.12: The *a posteriori* SPP estimates for the input given in Figure 5.11(b). In contrast to Figure 5.11(c) the estimators in this figure are capable of yielding low SPP in speech absence. However, the estimator in (a) does not indicate speech presence at the fricatives around $l = 63$ and $l = 192$ which results in higher speech distortions as compared to the proposed estimators in (b) and (c).



(a) Clean signal.

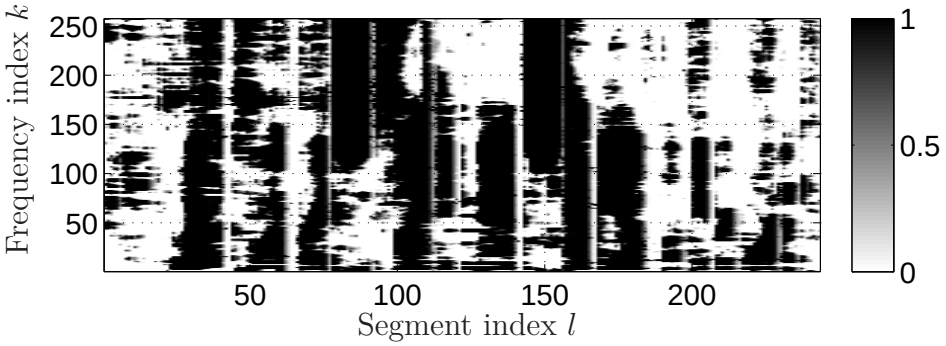


(b) Noisy signal.

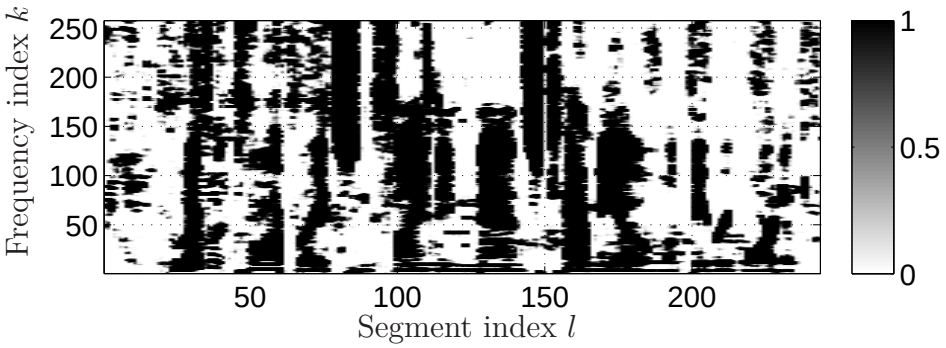


(c) *A posteriori* SPP according to [Malah *et al*, 1999].

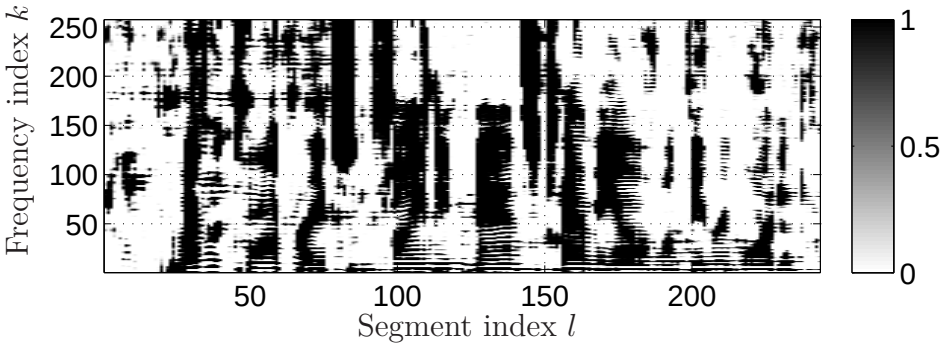
Figure 5.13: Clean speech (a), noisy speech (b), and the resulting *a posteriori* SPP estimate using [Malah *et al*, 1999] (c) for the sentence "Whoever cooperates in finding Nan's cameo will be rewarded" spoken by a male speaker disturbed by additive babble noise at 0 dB input segmental SNR. The signals in the spectrograms (a) and (b) have been pre-emphasized for a better visualization of high-frequency components. The estimator of [Malah *et al*, 1999] in (c) does not yield SPP estimates close to zero in speech absence.



(a) *A posteriori* SPP according to [Cohen and Berdugo, 2001].



(b) Frequency domain smoothing proposed in Section 5.4 [Gerkmann *et al*, 2008b].



(c) Cepstral domain smoothing proposed in Section 5.5 [Gerkmann *et al*, 2010].

Figure 5.14: The *a posteriori* SPP estimates for the input given in Figure 5.13(b). In contrast to Figure 5.13(c) the estimators in this figure are capable of yielding low SPP in speech absence. However, the estimators in (a) and (b) are not capable of resolving the spectral harmonics of the male speaker. This is clearly improved with the proposed approach in panel (c), resulting in less noise leakage without an increase in speech distortions.

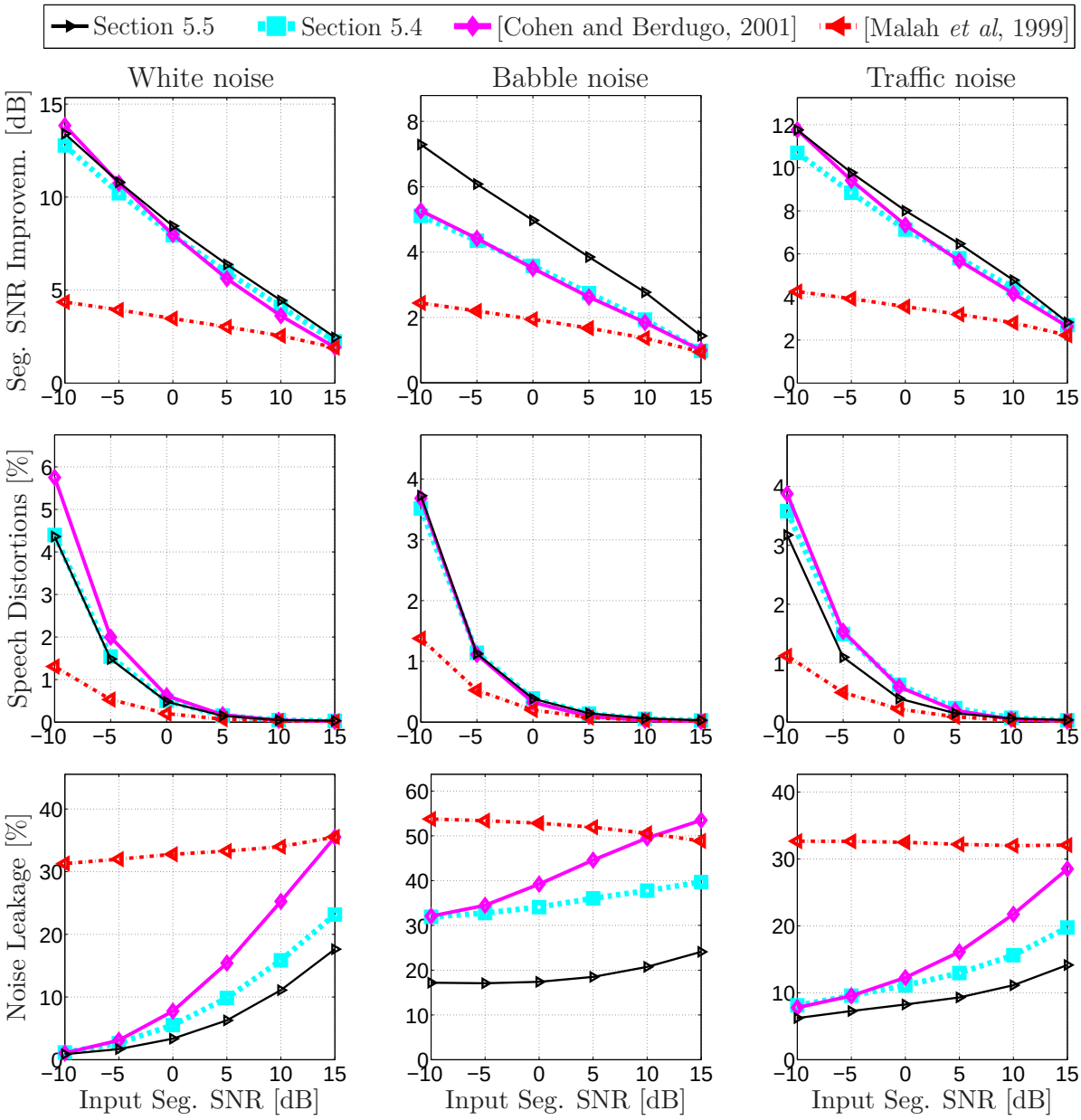


Figure 5.15: The average segmental SNR improvement (top), speech distortions (middle), and noise leakage (bottom) averaged over 320 TIMIT sentences for white Gaussian noise (left), babble noise (middle), and nonstationary traffic noise (right).

5.7 Conclusions

In this chapter, the theoretical basis for an *a posteriori* Speech Presence Probability (SPP) estimator based on a smoothed *a posteriori* Signal-to-Noise Ratio (SNR) is given. Smoothing the *a posteriori* SNR has the major benefit of reducing the variance of the SPP estimate. By interpreting the estimator as a detector, it is shown that this increases the estimation performance in terms of a lower false-alarm rate and a lower missed-hit rate. The perceptual benefits are less musical noise and less speech distortions when the SPP estimator is incorporated into a speech enhancement framework. In the cepstral domain *a priori* knowledge about which cepstral coefficients are likely to represent speech can be exploited to apply a selective temporal smoothing that reduces spectral outliers while preserving the speech spectral structure. Such a selective Temporal Cepstrum Smoothing (TCS) is shown to outperform the smoothing in the frequency domain proposed in [Cohen and Berdugo, 2001, Sørensen and Andersen, 2005] and Section 5.4. Further, while the approaches of [Cohen and Berdugo, 2001, Sørensen and Andersen, 2005] and Section 5.4 require to combine SPP estimates based on local and global spectral smoothing, this is not necessary with the proposed selective TCS.

The *a posteriori* SPP estimator is based on the ratio of the likelihoods of speech presence and speech absence, weighted by their prior probabilities. In state-of-the-art *a posteriori* SPP estimators the likelihood-ratio is usually based on an adaptively estimated *a priori* SNR estimate that takes very small values at time-frequency points where speech is absent (*e.g.* between the harmonics of voiced speech). We have shown that then the resulting *a posteriori* SPP estimate yields only the prior probabilities. Competing approaches attempt to mend this undesired behavior by adaptively estimating the speech presence priors.

However, we have argued that for speech presence probability estimation neither the *a priori* SNR nor the *a priori* SPP should be adapted, but reflect true prior knowledge. In particular, the *a priori* SNR should reflect the SNR that is expected when speech is present. To achieve this, an optimal fixed *a priori* SNR is used that minimizes the false-alarm and missed-hit rates. For the *a priori* SPP it is assumed that speech presence and absence are equally likely and set $P(\mathcal{H}_{1,k}) = 0.5$. Our modifications provide low *a posteriori* SPP estimates at time-frequency points where speech is absent, without the necessity for adaptively tracking the *a priori* SPP. Further, since the *a priori* SNR is not adaptively estimated, the proposed procedure enables a decoupling of the estimation of the speech presence probability and the estimation of the clean-speech spectral coefficients.

The proposed cepstral approach is shown to achieve a higher frequency resolution, considerably less noise leakage, and a higher or similar SNR improvement, while obtaining lower or similar speech distortions as compared to state-of-the-art estimators that yield

small values for the SPP in speech absence. At 0 dB SNR, the segmental SNR indicates a gain of approximately 1.5 dB as compared the competing algorithms in babble noise and a gain of approximately 0.5 dB in stationary white noise.

Chapter 6

Conclusions

In this thesis, Wiener filter based enhancement of noisy speech signals is addressed. The aim is to increase the signal-to-noise ratio improvement as compared to competing state-of-the art algorithms without increasing the amount of musical noise or distorting the speech spectral structure. The cepstral domain is shown to be well suited to reduce spectral outliers that yield musical noise while preserving the speech spectral structure.

In the cepstral domain speech is shown to be represented by few coefficients, thus enabling a selective modification of the speech related coefficients and the remaining coefficients. While the speech related coefficients are hardly modified, the remaining coefficients can be temporally smoothed, nulled, or replaced. Thus, spectral outliers that are represented by the remaining cepstral coefficients can be effectively reduced while the speech related cepstral coefficients are preserved.

To optimize the performance of cepstral modification techniques, it is important to understand the statistical properties of cepstral coefficients before and after modification. Thus, the statistical properties of cepstral coefficients and the logarithmic periodogram from χ-distributed spectral amplitudes and tapered spectral analysis windows are analyzed. In particular, explicit expressions for the mean and covariance matrix of the log-periodogram and cepstral coefficients are derived for spectrally uncorrelated, as well as spectrally correlated, χ-distributed spectral amplitudes. The spectral correlation introduced by tapered spectral analysis windows is shown to result in a decreasing cepstral variance for an increasing cepstral index. As the cepstral transformation includes a nonlinear compression, changing the variance of cepstral coefficients results in a bias in the spectral domain. As any of the proposed cepstral modification techniques — *i.e.* temporal cepstrum smoothing, cepstral nulling or cepstral replacement — results in a change of the average cepstral variance, a derivation of the bias for a given cepstral modification is of great interest. We have related the change of the average cepstral variance to the shape parameter of χ-distributed cepstral coefficients, and have shown that the shape parameter is increased for a cepstral variance reduction as obtained by temporal cepstrum

smoothing or cepstral nulling. As a result, the bias can be obtained as a function of the shape parameters before and after cepstral modification.

To determine the set of cepstral coefficients that represent the speech spectral structure, an estimate of the speech fundamental period is required. For uncorrelated spectral coefficients a maximum search in the upper cepstrum is shown to be the optimal fundamental period estimator in the maximum likelihood sense. Interestingly, if multiple microphones are present, the optimal estimator results in a maximum search on the sum of the microphone cepstra rather than a maximum search on the cepstrum of the output of a delay-and-sum beamformer. Further, the estimator is extended towards a maximum *a posteriori* fundamental period tracker that may further increase the performance.

Three applications for a temporal cepstrum smoothing are considered. First, temporal cepstrum smoothing is applied to the multiplicative gain function, then temporal cepstrum smoothing is used for the estimation of the *a priori* Signal-to-Noise Ratio (SNR). In Chapter 5 temporal cepstrum smoothing is applied for speech presence probability estimation. It is shown that temporal cepstrum smoothing reduces spectral outliers and results in a more natural sounding residual noise while keeping speech distortions low as compared to competing methods. In terms of instrumental measures, temporal cepstrum smoothing for *a priori* SNR estimation yields even better results than a smoothing of spectral gain functions. The advantage of a temporal cepstrum smoothing of spectral gain functions is its flexibility: it can be applied to any speech enhancement algorithm that uses multiplicative spectral gain functions, including binary masks for blind source separation. The computational complexity of cepstral smoothing approaches is dominated by two additional spectral transformations. However, the computational complexity can be greatly reduced if pruned Discrete Cosine Transforms (DCTs) are used for the cepstral transform and its inverse as proposed in Section 3.2.3.

One of the reasons why a temporal cepstrum smoothing for *a priori* SNR estimation performs better than a smoothing of spectral gain functions is that it is used in an earlier step of the speech enhancement framework. Motivated by this, it is proposed to apply a modification of cepstral coefficients in an even earlier step, namely the spectral noise power estimation. As an alternative to a temporal smoothing of the cepstrum, it is proposed to replace cepstral coefficients of the noise spectral power. When modifying the spectral noise power, extra care has to be taken that no speech information leaks into the noise power estimate as this would result in speech distortions. Therefore, as compared to the proposed temporal cepstrum smoothing algorithms, less cepstral coefficients can be modified in the noise power estimate. Thus, it is proposed to combine a spectral replacement in the noise power estimate with a replacement in the speech power estimate which is less sensitive to speech distortions. Instrumental measures indicate that the resulting framework produces comparable results as a temporal cepstrum smoothing for

a priori SNR estimation and thus in an increased performance as compared to competing methods without cepstral modification. The advantage of the cepstral replacement technique is that it works instantaneously in each time frame. For nonstationary noise, such as babble noise, this may be beneficial, as the background noise is not smeared over time. This effect was also confirmed by informal listening. The advantage of a temporal cepstrum smoothing is a much lower computational complexity, especially when pruned DCTs are used.

Finally, the problem of *a posteriori* speech presence probability estimation is addressed. While in state-of-the-art estimators the *a priori* speech presence probability and the *a priori* SNR are adaptively estimated, we have argued that for *a posteriori* speech presence probability the priors should not be adapted but represent true *a priori* knowledge. Further, it is proposed to smooth the *a posteriori* SNR in the frequency domain or, preferably, in the cepstral domain. Here, the determination of the shape parameter proposed in Section 2.3 is not only important to determine the bias, but also to derive the likelihoods of speech presence and absence. The estimator based on cepstral smoothing with optimally derived fixed priors and the proposed determination of the shape parameter is shown to clearly outperform state-of-the-art estimators in terms of speech distortions, noise leakage and segmental SNR improvement when multiplicatively applied to noisy speech.

Appendix A

Properties of the Cepstrum

A.1 Averaging independent χ^2-distributed random variables

In this section, we show that an unbiased averaging of independent χ^2-distributed random variables with shape parameter μ results in χ^2-distributed random variables with the same mean but an increase in the shape parameter that goes along with a decrease of the variance. Without loss of generality, we consider the special cases of an averaging of two and three random variables. The extension to an averaging of $L \in \mathbb{N}$ random variables, such as a moving average smoothing with a rectangular kernel of length L, is straightforward.

We want to derive the distribution of the random variable that results, when three independent χ^2-distributed random variable x, y, z are added and normalized, as $\bar{v} = (x + y + z)/3$. To achieve this, we first consider the addition of two χ^2-distributed random variables $w = x + y$. The two χ^2 distributions of x and y are given by

$$p_x(x) = \frac{1}{\Gamma(\mu)} \left(\frac{\mu}{\sigma^2}\right)^{\mu} x^{\mu-1} \exp\left(-\frac{\mu}{\sigma^2} x\right) \tag{A.1}$$

$$p_y(w-x) = \frac{1}{\Gamma(\mu)} \left(\frac{\mu}{\sigma^2}\right)^{\mu} (w-x)^{\mu-1} \exp\left(-\frac{\mu}{\sigma^2} (w-x)\right) \tag{A.2}$$

We assume that the two periodograms are χ^2-distributed with the same shape parameter μ, the same mean σ^2 and and the same variance σ^4/μ. In the context of a moving average smoothing, this corresponds to the assumption of a stationary process. Then, the distribution of the variable $w = x + y$ can be derived using [Papoulis and Pillai, 2002, (6–45)] as

$$\begin{aligned} p_w(w) &= \int_0^w p_x(x) p_y(w-x) \mathrm{d}x \\ &= \frac{1}{(\Gamma(\mu))^2} \left(\frac{\mu}{\sigma^2}\right)^{2\mu} \exp\left(-\frac{\mu}{\sigma^2} w\right) \int_0^w x^{\mu-1} (w-x)^{\mu-1} \, \mathrm{d}x \,. \end{aligned} \tag{A.3}$$

Using [Gradshteyn and Ryzhik, 2000, (3.191.1)] and [Gradshteyn and Ryzhik, 2000, (8.384.1)] we have

$$\int_0^w x^{\nu-1}(w-x)^{\mu-1}\mathrm{d}x = w^{\mu+\nu-1}\frac{\Gamma(\mu)\Gamma(\nu)}{\Gamma(\mu+\nu)}\,. \tag{A.4}$$

With (A.4) we can solve (A.3) and obtain

$$\begin{aligned} p_w(w) &= \frac{1}{\Gamma(2\mu)}\left(\frac{\mu}{\sigma^2}\right)^{2\mu} w^{2\mu-1}\exp\!\left(-\frac{\mu}{\sigma^2}w\right) \\ &= \frac{1}{\Gamma(2\mu)}\left(\frac{2\mu}{2\sigma^2}\right)^{2\mu} w^{2\mu-1}\exp\!\left(-\frac{2\mu}{2\sigma^2}w\right)\,. \end{aligned} \tag{A.5}$$

Comparing the χ^2-distribution (A.1) to (A.5) it can be seen that w is χ^2-distributed with shape parameter 2μ, mean $2\sigma^2$, and variance $2\sigma^4/\mu$. After normalizing w by 2, the smoothed periodogram $\bar{w} = \frac{1}{2}(x+y)$ is χ^2-distributed with $\mathrm{E}\{\bar{w}\} = \sigma^2$, shape parameter 2μ, and $\mathrm{var}\{\bar{w}\} = \sigma^4/(2\mu)$.

To obtain the distribution of $v = x+y+z = w+z$, we proceed with

$$p_z(z) = \frac{1}{\Gamma(\mu)}\left(\frac{\mu}{\sigma^2}\right)^{\mu} z^{\mu-1}\exp\!\left(-\frac{\mu}{\sigma^2}z\right) \tag{A.6}$$

$$p_w(v-z) = \frac{1}{\Gamma(2\mu)}\left(\frac{\mu}{\sigma^2}\right)^{2\mu} (v-z)^{2\mu-1}\exp\!\left(-\frac{\mu}{\sigma^2}(v-z)\right)\,. \tag{A.7}$$

Using (A.4), we obtain

$$\begin{aligned} p_v(v) &= \int_0^v p_z(z)p_w(v-z)\mathrm{d}z \\ &= \frac{1}{\Gamma(2\mu)\Gamma(\mu)}\left(\frac{\mu}{\sigma^2}\right)^{3\mu}\exp\!\left(-\frac{\mu}{\sigma^2}v\right)\int_0^v z^{\mu-1}\,(v-z)^{2\mu-1}\,\mathrm{d}z \\ &= \frac{1}{\Gamma(3\mu)}\left(\frac{\mu}{\sigma^2}\right)^{3\mu} v^{3\mu-1}\exp\!\left(-\frac{\mu}{\sigma^2}v\right) \\ &= \frac{1}{\Gamma(3\mu)}\left(\frac{3\mu}{3\sigma^2}\right)^{3\mu} v^{3\mu-1}\exp\!\left(-\frac{3\mu}{3\sigma^2}v\right)\,. \end{aligned} \tag{A.8}$$

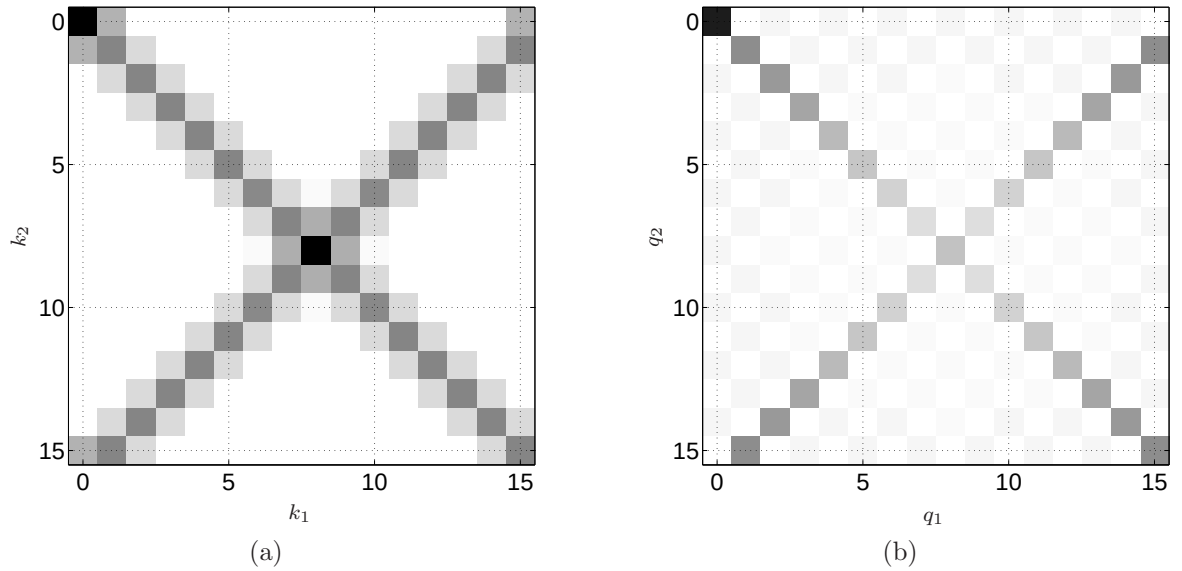
Thus, $v = x+y+z$ is χ^2-distributed with shape parameter 3μ, mean $3\sigma^2$, and variance $3\sigma^4/\mu$, while $\bar{v} = (x+y+z)/3$ is χ^2-distributed with shape parameter 3μ, mean σ^2, and variance $\sigma^4/(3\mu)$. The extension to an averaging of $L \in \mathbb{N}$ χ^2-distributed random variables with shape parameter μ and mean σ^2 is straightforward and results in a smoothed χ^2-distributed random variable with

- shape parameter $L\mu$,
- mean σ^2, and
- variance $\sigma^4/(L\mu)$.

A.2 Relation between the cepstral covariance and the log-periodogram

In this appendix, we show that the covariance of the cepstral coefficients can be obtained by taking a two dimensional discrete Fourier transform of the covariance of the log-periodogram. With the definition of the cepstrum (2.1) we obtain

$$\begin{aligned}
&\operatorname{cov}\{\phi_{q_1}, \phi_{q_2}\} \\
&= \mathrm{E}\{(\phi_{q_1} - \mathrm{E}\{\phi_{q_1}\})\,(\phi_{q_2} - \mathrm{E}\{\phi_{q_2}\})^*\} \\
&= \mathrm{E}\left\{ \frac{1}{N} \sum_{k_1=0}^{N-1} (\log P_{k_1} - \mathrm{E}\{\log P_{k_1}\})\, \mathrm{e}^{\mathrm{j}\frac{2\pi}{N} k_1 q_1} \cdot \frac{1}{N} \sum_{k_2=0}^{N-1} (\log P_{k_2} - \mathrm{E}\{\log P_{k_2}\})\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N} k_2 q_2} \right\} \\
&= \frac{1}{N^2} \sum_{k_2=0}^{N-1} \sum_{k_1=0}^{N-1} \operatorname{cov}\{\log P_{k_1}, \log P_{k_2}\}\, \mathrm{e}^{\mathrm{j}\frac{2\pi}{N} q_1 k_1} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N} q_2 k_2} \,. \qquad \text{(A.9)}
\end{aligned}$$



(a) (b)

Figure A.1: The covariance matrix of the log-periodogram $\operatorname{cov}\{\log(P_{k_1}), \log(P_{k_2})\}$ (a) and the cepstral coefficients $\operatorname{cov}\{\phi_{q_1}, \phi_{q_2}\}$ (b). The periodogram bins are obtained from a computer generated white Gaussian time domain signal, a Hann window with 50% overlap, and $N = 16$.

A.3 Cepstral covariance for correlated spectral coefficients

In this appendix, we derive an explicit expression for the covariance of cepstral coefficients, when the log-periodogram bins are correlated. The derived results hold for large N as usually used in speech enhancement applications. For large N, the covariance matrix of the log-periodogram can be approximated by a $N \times N$ symmetric circulant Toeplitz matrix [Gray, 2006] defined by the vector $[\kappa_0, \kappa_1, ..., \kappa_{N/2-1}, \kappa_{N/2}, \kappa_{N/2-1}, ..., \kappa_1]$, where we neglect the fact that for $k \in \{0, N/2\}$ the variance of the log-periodogram is larger than κ_0, as we have less degrees of freedom than for $k \notin \{0, N/2\}$. The covariance of the cepstral coefficients is obtained by taking a two dimensional discrete Fourier transform, as presented in Appendix A.2. As in general the spectral covariance κ_m introduced by tapered spectral analysis windows rapidly decreases with increasing m, we assume that $\kappa_m = 0$ for $m > M$ and $M \ll N/2 + 1$. Then, the covariance of cepstral coefficients results in

$$
\begin{aligned}
&\operatorname{cov}\{\phi_{q_1}, \phi_{q_2}\} \\
=&\frac{1}{N^2} \sum_{k_2=0}^{N-1} \Bigg(2\kappa_0 \cos\left(\frac{2\pi}{N} q_1 k_2\right) \\
&+ \sum_{m=1}^{M} 2\kappa_m \Bigg(\cos\left(\frac{2\pi}{N} q_1 \Big(k_2 - m\Big)\right) + \cos\left(\frac{2\pi}{N} q_1 \Big(k_2 + m\Big)\right) \Bigg) \Bigg) \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N} q_2 k_2} \\
=&\frac{1}{N^2} \Bigg(\kappa_0 \left(\sum_{k_2=0}^{N-1} \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}(q_1 - q_2)k_2} + \sum_{k_2=0}^{N-1} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}(q_1 + q_2)k_2} \right) \\
&+ \sum_{m=1}^{M} \kappa_m \Bigg(\mathrm{e}^{-\mathrm{j}\frac{2\pi}{N} q_1 m} \sum_{k_2=0}^{N-1} \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}(q_1 - q_2)k_2} + \mathrm{e}^{\mathrm{j}\frac{2\pi}{N} q_1 m} \sum_{k_2=0}^{N-1} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}(q_1 + q_2)k_2} \\
&\qquad + \mathrm{e}^{\mathrm{j}\frac{2\pi}{N} q_1 m} \sum_{k_2=0}^{N-1} \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}(q_1 - q_2)k_2} + \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N} q_1 m} \sum_{k_2=0}^{N-1} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}(q_1 + q_2)k_2} \Bigg) \Bigg) \\
=&\begin{cases} \frac{1}{N} \left(\kappa_0 + 2\sum_{m=1}^{M} \kappa_m \cos\left(\frac{2\pi}{N} q_1 m\right) \right) & , (q_1 = q_2 \neq 0, N/2) \text{ OR } (q_1 + q_2 = N) \\ \frac{2}{N} \left(\kappa_0 + 2\sum_{m=1}^{M} \kappa_m \cos\left(\frac{2\pi}{N} q_1 m\right) \right) & , q_1 = q_2 = 0, N/2 \\ 0 & , \text{ else.} \end{cases}
\end{aligned}
\tag{A.10}
$$

Note that (A.10) is the result for the full symmetric cepstrum $q \in \{0, ..., N-1\}$, while in (2.20) the solution for the lower symmetric part $q \in \{0, ..., N/2\}$ is given.

In Figure A.1 the covariance matrices of the log-periodogram and the cepstral coefficients are illustrated. There, the periodogram bins are obtained from a computer generated

white Gaussian time domain signal. The spectral analysis (1.1) is obtained with a Hann spectral analysis window w_n and $N = 16$. As N is relatively small, a slight correlation may be observed in Figure A.1(b) when both q_1 and q_2 are even, or both q_1 and q_2 are odd. These correlations arise from the fact, that for $k \in \{0, N/2\}$ the variance of the log-periodogram is larger than κ_0, which we neglected in the derivation of (A.10). However, the resulting correlations decrease with $1/N^2$ [Ephraim and Rahim, 1999]. For segment sizes N, as usually used in speech processing, these correlations are insignificant, *i.e.* the cepstral coefficients are asymptotically uncorrelated for large N.

A.4 Spectral correlation for a Hann window

In this appendix we derive the correlations ρ_1 and ρ_2 for a Hann window, with ρ_m defined in (2.19).

The multiplication of the time domain signal with the window function w_n in (1.1) results in a convolution of the uncorrelated spectral coefficients U_k with the Fourier domain representation of the window function $W_k = \mathrm{DFT}\{w_n\}$, *i.e.* $S_k = U_k * W_k$, where the asterisk denotes convolution and $\mathrm{DFT}\{\cdot\}$ the discrete Fourier transform. For a normalized discrete Hann window we obtain the correlated frequency coefficients

$$S_k = -\sqrt{\frac{1}{6}}U_{k-1} + \sqrt{\frac{2}{3}}U_k - \sqrt{\frac{1}{6}}U_{k+1}\,. \tag{A.11}$$

Because U_k is spectrally uncorrelated, with $\sigma^2_{k-1} \approx \sigma^2_k \approx \sigma^2_{k+1}$ and $\mathrm{E}\{|U_k|^2\} = \sigma^2_k$ we have

$$\mathrm{E}\Big\{|S_k|^2\Big\} = \sigma^2_k\,. \tag{A.12}$$

For the covariances we obtain

$$\mathrm{E}\Big\{S_k S^*_{k+1}\Big\} = -\frac{1}{3}\sigma^2_k - \frac{1}{3}\sigma^2_{k+1}\,, \tag{A.13}$$

$$\mathrm{E}\Big\{S_k S^*_{k+2}\Big\} = -\frac{1}{6}\sigma^2_{k+1}\,. \tag{A.14}$$

Thus, with $\sigma^2_{k-1} \approx \sigma^2_k \approx \sigma^2_{k+1}$ and (2.19) we have $\rho_1 = 2/3$ and $\rho_2 = 1/6$.

A.5 Variance after smoothing

In this appendix we relate the variance of recursively and moving average smoothed random variables to the variance of the nonsmoothed variable.

A.5.1 Variance after recursive smoothing

The recursive smoothing of a variable s is given by:

$$\bar{s}(l) = \alpha\bar{s}(l-1) + (1-\alpha)s(l)\,. \tag{A.15}$$

For stationary processes, for the first moment follows

$$\begin{aligned} \mathrm{E}\{\bar{s}\} &= \alpha\mathrm{E}\{\bar{s}\} + (1-\alpha)\mathrm{E}\{s\} \\ (1-\alpha)\mathrm{E}\{\bar{s}\} &= (1-\alpha)\mathrm{E}\{s\} \\ \mathrm{E}\{\bar{s}\} &= \mathrm{E}\{s\}\,. \end{aligned} \tag{A.16}$$

For uncorrelated successive $s(l)$, the smoothed quantity $\bar{s}(l-1)$ and $s(l)$ are uncorrelated. Then, the second moment is given by:

$$\begin{aligned} \mathrm{E}\left\{\bar{s}^2\right\} &= \alpha^2\mathrm{E}\left\{\bar{s}^2\right\} + (1-\alpha)^2\mathrm{E}\left\{s^2\right\} + 2\alpha(1-\alpha)(\mathrm{E}\{s\})^2 \\ &= \frac{1-\alpha}{1+\alpha}\mathrm{E}\left\{s^2\right\} + 2\frac{\alpha}{1+\alpha}(\mathrm{E}\{s\})^2 \end{aligned} \tag{A.17}$$

Thus, for the variance we obtain

$$\begin{aligned} \mathrm{var}\{\bar{s}\} &= \mathrm{E}\left\{\bar{s}^2\right\} - (\mathrm{E}\{\bar{s}\})^2 \\ &= \frac{1-\alpha}{1+\alpha}\mathrm{E}\left\{s^2\right\} - (\mathrm{E}\{s\})^2\left(1 - 2\frac{\alpha}{1+\alpha}\right) \\ &= \frac{1-\alpha}{1+\alpha}\mathrm{var}\{s\}\,. \end{aligned} \tag{A.18}$$

A.5.2 Variance after moving average smoothing

A moving average is obtained by taking the average over L frames, as

$$\bar{s}(l) = \frac{1}{L}\sum_{\ell=0}^{L-1} s(l-\ell)\,. \tag{A.19}$$

For stationary processes, the first moment is given by

$$\begin{aligned} \mathrm{E}\{\bar{s}\} &= \frac{1}{L}\sum_{\ell=0}^{L-1}\mathrm{E}\{s(l-\ell)\} \\ &= \mathrm{E}\{s\}\,. \end{aligned} \tag{A.20}$$

For uncorrelated successive $s(l)$, with [Gradshteyn and Ryzhik, 2000, (0.121.1)] the second moment is given by:

$$\begin{aligned} \mathrm{E}\left\{\bar{s}^2\right\} &= \frac{1}{L^2}\mathrm{E}\left\{\left(\sum_{\ell=0}^{L-1} s(l-\ell)\right)^2\right\} \\ &= \frac{1}{L^2}\left(L\,\mathrm{E}\left\{s^2\right\} + 2\frac{L(L-1)}{2}\,\left(\mathrm{E}\{s\}\right)^2\right) \\ &= \frac{1}{L}\mathrm{E}\left\{s^2\right\} + (1-\frac{1}{L})\,\left(\mathrm{E}\{s\}\right)^2 \end{aligned} \tag{A.21}$$

Thus, for the variance we obtain

$$\begin{aligned} \mathrm{var}\{\bar{s}\} &= \mathrm{E}\left\{\bar{s}^2\right\} - \left(\mathrm{E}\{\bar{s}\}\right)^2 \\ &= \frac{1}{L}\mathrm{E}\left\{s^2\right\} - \frac{1}{L}\,\left(\mathrm{E}\{s\}\right)^2 \\ &= \frac{1}{L}\,\mathrm{var}\{s\}\,. \end{aligned} \tag{A.22}$$

A.5.3 Relation between recursive smoothing and moving average smoothing

From (A.18) and (A.22) we can compute the recursive smoothing constant that results in the same variance reduction as a moving average smoothing for uncorrelated and stationary processes:

$$L = \frac{\mathrm{var}\{s\}}{\mathrm{var}\{\bar{s}\}} = \frac{1+\alpha}{1-\alpha} \ . \tag{A.23}$$

Vice versa we find

$$\alpha = \frac{L-1}{L+1} \ . \tag{A.24}$$

Appendix B

Temporal Cepstrum Smoothing for Speech Enhancement

B.1 Derivation of the maximum likelihood a priori SNR

In this appendix we derive the Maximum Likelihood (ML) *a priori* Signal-to-Noise Ratio (SNR) estimate for a χ^2-distributed *a posteriori* SNR. The ML *a priori* SNR estimate is obtained by maximizing the likelihood function, as

$$\xi_k^{\mathrm{ml}} = \arg\max_{\xi_k} p\Big(\gamma_k \;\Big|\; \xi_k\Big)\,, \tag{B.1}$$

with the likelihood given by

$$p\Big(\gamma_k \;\Big|\; \xi_k\Big) = \frac{1}{\Gamma(\mu)} \left(\frac{\mu}{1+\xi_k}\right)^{\mu} \gamma_k^{\mu-1} \exp\!\left(-\mu\,\frac{\gamma_k}{1+\xi_k}\right). \tag{B.2}$$

To find the maximum of the likelihood, we set its first derivative with respect to ξ_k to zero

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\xi_k}\, p\Big(\gamma_k \;\Big|\; \xi_k\Big) \stackrel{!}{=} 0 \\
&= \frac{1}{\Gamma(\mu)} \mu^{\mu}\, \gamma_k^{\mu-1} \exp\!\left(-\mu\,\frac{\gamma_k}{1+\xi_k}\right) \Big(-\mu(1+\xi_k)^{-\mu-1} + (1+\xi_k)^{-\mu}\mu\gamma_k(1+\xi_k)^{-2}\Big)\,.
\end{aligned} \tag{B.3}$$

We then obtain

$$\mu(1+\xi_k)^{-\mu-1} = \mu\gamma_k(1+\xi_k)^{-\mu-2} \tag{B.4}$$

$$1 = \gamma_k(1+\xi_k)^{-1}\,, \tag{B.5}$$

and thus finally

$$\xi_k^{\mathrm{ml}} = \gamma_k - 1\,, \tag{B.6}$$

which is a maximum, as the χ^2-distribution is unimodal for $\gamma_k \geq 0$.

Bibliography

[Andrianakis and White, 2006] Ioannis Andrianakis and Paul R. White. "MMSE speech spectral amplitude estimators with Chi and Gamma speech priors." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1068–1071, May 2006.

[Andrianakis and White, 2009] Ioannis Andrianakis and Paul R. White. "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors." *ELSEVIER Speech Communication*, volume 51, no. 1, pages 1–14, January 2009.

[Bogert *et al*, 1963] Bruce P. Bogert, Michael J. R. Healy, and John W. Tukey. "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking." In Murray Rosenblatt, editor, *Proceedings of the Symposium on Time Series Analysis*, pages 209–243. Wiley, New York, NY, USA, 1963.

[Brandt and Bitzer, 2009] Matthias Brandt and Joerg Bitzer. "Optimal spectral smoothing in short-time spectral attenuation (STSA) algorithms: Results of objective measures and listening tests." *EURASIP European Signal Processing Conference (EUSIPCO)*, pages 199–203, August 2009.

[Breithaupt, 2008] Colin Breithaupt. *Noise Reduction Algorithms for Speech Communications – Statistical Analysis and Improved Estimation Procedures.* Ph.D. thesis, Ruhr-Universität Bochum, Bochum, Germany, 2008.

[Breithaupt and Gerkmann, 2007] Colin Breithaupt and Timo Gerkmann. "Cepstral smoothing: Audio examples." December 2007. `http://www2.ika.rub.de/audioexamples/csmooth/csmooth.html`.

[Breithaupt *et al*, 2007] Colin Breithaupt, Timo Gerkmann, and Rainer Martin. "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise." *IEEE Signal Processing Letters*, volume 14, no. 12, pages 1036–1039, December 2007.

[Breithaupt *et al*, 2008a] Colin Breithaupt, Timo Gerkmann, and Rainer Martin. "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4897–4900, April 2008a.

[Breithaupt *et al*, 2008b] Colin Breithaupt, Martin Krawczyk, and Rainer Martin. "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4037–4040, April 2008b.

[Breithaupt and Martin, 2010] Colin Breithaupt and Rainer Martin. "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions." to appear in *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[Brillinger, 1981] David R. Brillinger. *Time Series: Data Analysis and Theory*. Holden-Day, San Francisco, CA, USA, 1981.

[Cappé, 1994] Olivier Cappé. "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor." *IEEE Transactions on Speech and Audio Processing*, volume 2, no. 2, pages 345–349, April 1994.

[Cheveigné and Kawahara, 2002] Alain de Cheveigné and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music." *Journal of the Acoustical Society of America*, volume 111, no. 4, pages 1917–1930, April 2002.

[Cohen, 2003] Israel Cohen. "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging." *IEEE Transactions on Speech and Audio Processing*, volume 11, no. 5, pages 466–475, September 2003.

[Cohen and Berdugo, 2001] Israel Cohen and Baruch Berdugo. "Speech enhancement for non-stationary noise environments." *ELSEVIER Signal Processing*, volume 81, no. 11, pages 2403–2418, November 2001.

[Cooley and Tukey, 1965] James W. Cooley and John W. Tukey. "An algorithm for the machine calculation of complex Fourier series." *Mathematics of Computation*, volume 19, no. 90, pages 297–301, 1965.

[Deller *et al*, 1993] John R. Deller, John H. L. Hansen, and John G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, New York, NY, USA, 1993.

[Dreiseitel and Schmidt, 2006] Pia Dreiseitel and Gerhard Schmidt. "Evaluation of algorithms for speech enhancement." In Eberhard Hänsler and Gerhard Schmidt, editors, *Topics in Acoustic Echo and Noise Control*. Springer Verlag, Berlin, Heidelberg, Germany, 2006.

[Droppo and Acero, 1998] James Droppo and Alex Acero. "Maximum a posteriori pitch tracking." *International Conference on Spoken Language Processing (ICSLP)*, pages 943–946, December 1998.

[Ephraim and Cohen, 2006] Yariv Ephraim and Israel Cohen. "Recent advancements in speech enhancement." In Richard C. Dorf, editor, *The Electrical*

Engineering Handbook. CRC Press, Taylor & Francis Group, Boca Raton, FL, USA, third edition, 2006.

[Ephraim and Malah, 1984] Yariv Ephraim and David Malah. "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator." *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 32, no. 6, pages 1109–1121, December 1984.

[Ephraim and Malah, 1985] Yariv Ephraim and David Malah. "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator." *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 33, no. 2, pages 443–445, April 1985.

[Ephraim and Rahim, 1999] Yariv Ephraim and Mazin Rahim. "On second-order statistics and linear estimation of cepstral coefficients." *IEEE Transactions on Speech and Audio Processing*, volume 7, no. 2, pages 162–176, March 1999.

[Ephraim and Roberts, 2005] Yariv Ephraim and William J. J. Roberts. "On second-order statistics of log-periodogram with correlated components." *IEEE Signal Processing Letters*, volume 12, no. 9, pages 625–628, September 2005.

[Erkelens *et al*, 2007a] Jan Erkelens, Jesper Jensen, and Richard Heusdens. "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria." *ELSEVIER Speech Communication*, volume 49, no. 7–8, pages 530–541, July 2007a.

[Erkelens *et al*, 2007b] Jan S. Erkelens, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 15, no. 6, pages 1741–1752, August 2007b.

[Esch *et al*, 2010] Thomas Esch, Florian Heese, Bernd Geiser, and Peter Vary. "Wideband noise suppression supported by artificial bandwidth extension techniques." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4790–4793, March 2010.

[Esch and Vary, 2009] Thomas Esch and Peter Vary. "Efficient musical noise suppression for speech enhancement systems." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4409–4412, 2009.

[Fingscheidt *et al*, 2005] Tim Fingscheidt, Christophe Beaugeant, and Suhadi Suhadi. "Overcoming the statistical independence assumption w.r.t. frequency in speech enhancement." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1081–1084, 2005.

[Flego, 2006] Federico Flego. *Fundamental Frequency Estimation Techniques for Multi-Microphone Speech Input*. Ph.D. thesis, University of Trento, Italy, March 2006.

[Garofolo, 1988] John S. Garofolo. "DARPA TIMIT acoustic-phonetic speech database." *National Institute of Standards and Technology (NIST)*, 1988.

[Gerkmann *et al*, 2008a] Timo Gerkmann, Colin Breithaupt, and Rainer Martin. "Bias compensation for cepstro-temporal smoothing of spectral filter gains." *ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, October 2008a.

[Gerkmann *et al*, 2008b] Timo Gerkmann, Colin Breithaupt, and Rainer Martin. "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 16, no. 5, pages 910–919, July 2008b.

[Gerkmann *et al*, 2010] Timo Gerkmann, Martin Krawczyk, and Rainer Martin. "Speech presence probability estimation based on temporal cepstrum smoothing." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4254–4257, March 2010.

[Gerkmann and Martin, 2006] Timo Gerkmann and Rainer Martin. "Soft decision combining for dual channel noise reduction." *ISCA Interspeech – Conference on Speech Communication and Technology*, pages 2134–2137, September 2006.

[Gerkmann and Martin, 2009] Timo Gerkmann and Rainer Martin. "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling." *IEEE Transactions on Signal Processing*, volume 57, no. 11, pages 4165–4174, November 2009.

[Gerkmann and Martin, 2010a] Timo Gerkmann and Rainer Martin. "Cepstral smoothing with reduced computational complexity." *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, October 2010a. Accepted for publication.

[Gerkmann and Martin, 2010b] Timo Gerkmann and Rainer Martin. "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances." *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, August 2010b. Submitted.

[Gerkmann *et al*, 2009] Timo Gerkmann, Rainer Martin, and Derya Dalga. "Multi-microphone maximum a posteriori fundamental frequency estimation in the cepstral domain." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4505–4508, April 2009.

[Goh *et al*, 1998] Zenton Goh, Kah-Chye Tan, and Bernard T. G. Tan. "Postprocessing method for suppressing musical noise generated by spectral subtraction." *IEEE Transactions on Speech and Audio Processing*, volume 6, no. 3, pages 287–292, May 1998.

[Gradshteyn and Ryzhik, 2000] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals Series and Products*. Academic Press, San Diego, CA, USA, 6th edition, 2000. Alan Jeffrey and Daniel Zwillinger, editors.

[Gray, Jr, 1974] Augustine H. Gray, Jr. "Log spectra of Gaussian signals." *Journal of the Acoustical Society of America*, volume 55, no. 5, pages 1028–1033, May 1974.

[Gray, 2006] Robert M. Gray. *Toeplitz and Circulant Matrices: A review.* now Publisher Inc., Hanover, MA, USA, 2006.

[Gustafsson *et al*, 2001] Harald Gustafsson, Sven Nordholm, and Ingvar Claesson. "Spectral subtraction using reduced delay convolution and adaptive averaging." *IEEE Transactions on Speech and Audio Processing*, volume 9, no. 8, pages 799–807, November 2001.

[Habets *et al*, 2008] Emanuel A. P. Habets, Israel Cohen, and Sharon Gannot. "Generating nonstationary multisensor signals under a spatial coherence constraint." *Journal of the Acoustical Society of America*, volume 124, no. 5, pages 2911–2917, November 2008.

[Hendriks *et al*, 2010] Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. "MMSE based noise PSD tracking with low complexity." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4266–4269, March 2010.

[Hendriks *et al*, 2008] Richard C. Hendriks, Jesper Jensen, and Richard Heusdens. "Noise tracking using DFT domain subspace decompositions." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 16, no. 3, pages 541–553, March 2008.

[Hendriks and Martin, 2007] Richard C. Hendriks and Rainer Martin. "MAP estimators for speech enhancement under normal and Rayleigh inverse Gaussian distributions." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 15, no. 3, pages 918–927, March 2007.

[Hess, 1983] Wolfgang J. Hess. *Pitch Determination of Speech Signals: Algorithms and Devices.* Springer Verlag, Berlin, Germany, 1983.

[Hu and Wang, 2004] Guoning Hu and DeLiang Wang. "Monaural speech segregation based on pitch tracking and amplitude modulation." *IEEE Transactions on Neural Networks*, volume 15, no. 5, pages 1135–1150, September 2004.

[Hyvärinen *et al*, 2001] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis.* John Wiley & Sons, Inc., New York, NY, USA, 2001.

[ITU-T, 2001] ITU-T. "Perceptual evaluation of speech quality (PESQ)." *ITU-T Recommendation P.862*, 2001.

[Jan *et al*, 2009] Tariqullah Jan, Wenwu Wang, and DeLiang Wang. "A multistage approach for blind separation of convolutive speech mixtures." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1713–1716, April 2009.

[Jax and Vary, 2003] Peter Jax and Peter Vary. "On artificial bandwidth extension of telephone speech." *ELSEVIER Signal Processing*, volume 83, no. 8, pages 1707–1719, August 2003.

[Joarder, 2009] Anwar H. Joarder. "Moments of the product and ratio of two correlated chi-square variables." *Springer Statistical Papers*, volume 50, no. 3, pages 581–592, June 2009.

[Larsen and Aarts, 2004] Erik Larsen and Ronald M. Aarts, editors. *Audio Bandwidth Extension*. John Wiley & Sons, Chichester, West Sussex, UK, November 2004.

[Linhard and Haulick, 1999] Klaus Linhard and Tim Haulick. "Noise subtraction with parametric recursive gain curves." *ISCA Eurospeech – European Conference on Speech Communication and Technology*, pages 2611–2614, September 1999.

[Lo and Cham, 1996] Kwok-Tung Lo and Wai-Kuen Cham. "Analysis of pruning in fast cosine transform." *IEEE Transactions on Signal Processing*, volume 44, no. 3, pages 714–717, March 1996.

[Loizou, 2005] Philipos C. Loizou. "Speech enhancement based in perceptually motivated bayesian estimators of the magnitude spectrum." *IEEE Transactions on Speech and Audio Processing*, volume 13, no. 5, pages 857–869, September 2005.

[Loizou, 2007] Philipos C. Loizou. *Speech Enhancement - Theory and Practice*. CRC Press, Taylor & Francis Group, Boca Raton, FL, USA, 2007.

[Lotter, 2004] Thomas Lotter. *Single and Multimicrophone Speech Enhancement for Hearing Aids*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, 2004.

[Lotter and Vary, 2005] Thomas Lotter and Peter Vary. "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model." *EURASIP Journal of Applied Signal Processing*, volume 2005, no. 7, pages 1110–1126, January 2005.

[Madhu *et al*, 2008] Nilesh Madhu, Colin Breithaupt, and Rainer Martin. "Temporal smoothing of spectral masks in the cepstral domain for speech separation." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 45–48, April 2008.

[Malah *et al*, 1999] David Malah, Richard Cox, and Anthony Accardi. "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 789–792, March 1999.

[Markel and Gray, 1976] John D. Markel and Augustine H. Gray, Jr. *Linear prediction of Speech*. Springer Verlag, Berlin, Germany, 1976.

[Martin, 2001] Rainer Martin. "Noise power spectral density estimation based on optimal smoothing and minimum statistics." *IEEE Transactions on Speech and Audio Processing*, volume 9, no. 5, pages 504–512, July 2001.

[Martin, 2002] Rainer Martin. "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 253–256, May 2002.

[Martin, 2005] Rainer Martin. "Speech enhancement based on minimum mean-square error estimation and supergaussian priors." *IEEE Transactions on Speech and Audio Processing*, volume 13, no. 5, pages 845–856, September 2005.

[Martin, 2006] Rainer Martin. "Bias compensation methods for minimum statistics noise power spectral density estimation." *ELSEVIER Signal Processing*, volume 86, no. 6, pages 1215–1229, June 2006.

[Martin and Lotter, 2001] Rainer Martin and Thomas Lotter. "Optimal recursive smoothing of non-stationary periodograms." *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 167–170, September 2001.

[Mauler *et al*, 2008] Dirk Mauler, Timo Gerkmann, and Rainer Martin. "An analysis of quefrency selective temporal smoothing of the cepstrum in speech enhancement." *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, WA, USA, September 2008.

[Mauler and Martin, 2007] Dirk Mauler and Rainer Martin. "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement." *EURASIP European Signal Processing Conference (EUSIPCO)*, pages 222–227, September 2007.

[Mauler and Martin, 2010] Dirk Mauler and Rainer Martin. "Optimization of switchable windows for low-delay spectral analysis-synthesis." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4718–4721, March 2010.

[McAulay and Malpass, 1980] Robert J. McAulay and Marilyn L. Malpass. "Speech enhancement using a soft-decision noise suppression filter." *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 28, no. 2, pages 137–145, April 1980.

[Meyer, Accessed 2006] Georg Meyer. "Keele pitch database." *University of Liverpool, School of Psychology*, Accessed 2006. `http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html`.

[Nadarajah, 2009] Saralees Nadarajah. "Comment on the paper by A. H. Joarder." *Springer Statistical Papers*, volume 50, no. 2, pages 441–443, March 2009.

[Noll, 1967] A. Michael Noll. "Cepstrum pitch estimation." *Journal of the Acoustical Society of America*, volume 41, pages 293–309, February 1967.

[Oppenheim and Schafer, 1975] Alan V. Oppenheim and Ronald W. Schafer. *Digital Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, USA, 1975.

[Oppenheim and Schafer, 2004] Alan V. Oppenheim and Ronald W. Schafer. "From frequency to quefrency: A history of the cepstrum." *IEEE Signal Processing Magazine*, volume 21, no. 5, pages 95–106, September 2004.

[Papoulis and Pillai, 2002] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, NY, USA, fourth edition, 2002.

[Porter and Boll, 1984] Jack E. Porter and Steven F. Boll. "Optimal estimators for spectral restoration of noisy speech." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 18A.2.1–18A.2.4, 1984.

[Quackenbush *et al*, 1988] Schuyler R. Quackenbush, Thomas P. Barnwell, III, and Mark A. Clements. *Objective Measures of Speech Quality*. Prentice Hall, Englewood Cliffs, NJ, USA, 1988.

[Rosca *et al*, 2006] Justinian Rosca, Timo Gerkmann, and Doru-Cristian Balcan. "Statistical inference of missing speech data in the ICA domain." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 617–620, May 2006.

[Schafer and Rabiner, 1969] Ronald W. Schafer and Lawrence R. Rabiner. "System for automatic formant analysis of voiced speech." *Journal of the Acoustical Society of America*, volume 47, no. 2, pages 634–648, February 1969.

[Sorensen *et al*, 1987] Henrik V. Sorensen, Douglas L. Jones, Michael T. Heideman, and C. Sidney Burrus. "Real-valued fast Fourier transform algorithms." *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 35, no. 6, pages 849–863, June 1987.

[Sørensen and Andersen, 2005] Karsten V. Sørensen and Søren V. Andersen. "Speech enhancement with natural sounding residual noise based on connected time-frequency speech presence regions." *EURASIP Journal on Applied Signal Processing*, volume 2005, no. 18, pages 2954–2964, 2005.

[Srinivasan *et al*, 2006] Sriram Srinivasan, Jonas Samuelsson, and W. Bastiaan Kleijn. "Codebook driven short-term predictor parameter estimation for speech enhancement." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 14, no. 1, pages 163–176, January 2006.

[Srinivasan *et al*, 2007] Sriram Srinivasan, Jonas Samuelsson, and W. Bastiaan Kleijn. "Codebook-based bayesian speech enhancement for nonstationary

environments." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 15, pages 441–452, February 2007.

[Stoica and Sandgren, 2006] Petre Stoica and Niclas Sandgren. "Smoothed nonparametric spectral estimation via cepstrum thresholding." *IEEE Signal Processing Magazine*, volume 23, no. 6, pages 34–45, November 2006.

[Stoica and Sandgren, 2007] Petre Stoica and Niclas Sandgren. "Total-variance reduction via thresholding: Application to cepstral analysis." *IEEE Transactions on Signal Processing*, volume 55, no. 1, pages 66–72, January 2007.

[Tabrikian *et al*, 2004] Joseph Tabrikian, Shlomo Dubnov, and Yulya Dickalov. "Maximum a posteriori probability pitch tracking in noisy environments using harmonic model." *IEEE Transactions on Speech and Audio Processing*, volume 12, no. 1, pages 76–87, January 2004.

[Tilp, 2002] Jan Tilp. *Verfahren zur Verbesserung gestörter Sprachsignale unter Berücksichtigung der Grundfrequenz stimmhafter Laute*. Ph.D. thesis, Universität Darmstadt, Darmstadt, Germany, July 2002.

[Van Trees, 1968] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory. Part I.* John Wiley & Sons, New York, NY, USA, 1968.

[Vary and Martin, 2006] Peter Vary and Rainer Martin. *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley & Sons, Chichester, West Sussex, UK, 2006.

[Wang, 1984] Zhongde Wang. "Fast algorithms for the discrete W transform and for the discrete Fourier transform." *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 32, no. 4, pages 803–812, August 1984.

[Wang, 1991] Zhongde Wang. "Pruning the fast discrete cosine transform." *IEEE Transactions on Communications*, volume 39, no. 5, pages 640–643, May 1991.

[Wolfe and Godsill, 2001] Patrick J. Wolfe and Simon J. Godsill. "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement." *IEEE Workshop on Statistical Signal Processing*, pages 496–499, August 2001.

[Yegnanarayana and Murty, 2009] Bayya Yegnanarayana and K. Sri Rama Murty. "Event-based instantaneous fundamental frequency estimation from speech signals." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 17, no. 4, pages 614–624, May 2009.

[You *et al*, 2005] Chang Huai You, Soo Ngee Koh, and Susanto Rahardja. "β-order MMSE spectral amplitude estimation for speech enhancement." *IEEE Transactions on Speech and Audio Processing*, volume 13, no. 4, pages 475–486, July 2005.

[Yu and Hansen, 2009] Tao Yu and John H. L. Hansen. "A speech presence microphone array beamformer using model based speech presence probability estima-

tion." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 213–216, April 2009.