Technische Universität Dresden

Robuste Spracherkennung unter raumakustischen Umgebungsbedingungen

Rico Petrick

von der Fakultät Elektrotechnik und Informationstechnik der Technischen Universität Dresden

zur Erlangung des akademischen Grades eines

Doktoringenieurs

(Dr.-Ing.)

vorgelegte Dissertation

Gutachter:	Prof. Dr.-Ing. habil. Rüdiger Hoffmann (TU Dresden) Prof. Dr. Masashi Unoki (Japan Advanced Institute of Science and Technology)
Tag der Einreichung:	30. April 2009
Tag der Verteidigung:	25. September 2009

Abstract

Automatic speech recognition (ASR) systems used in real-world indoor scenarios suffer from performance degradation if noise and reverberation conditions differ from the training conditions of the recognizer. This thesis deals with the problem of room reverberation as a cause of distortion in ASR systems. The background of this research is the design of practical command and control applications, such as a voice controlled light switch in rooms or similar applications. Therefore, the design aims to incorporate several restricting working conditions for the recognizer and still achieve a high level of robustness. One of those design restrictions is the minimisation of computational complexity to allow the practical implementation on an embedded processor.

One chapter comprehensively describes the room acoustic environment, including the behavior of the sound field in rooms. It addresses the speaker room microphone (SRM) system which is expressed in the time domain as the room impulse response (RIR). The convolution of the RIR with the clean speech signal yields the reverberant signal at the microphone.

A thorough analysis proposes that the degree of the distortion caused by reverberation is dependent on two parameters, the reverberation time T_{60} and the speaker-to-microphone distance (SMD). To evaluate the dependency of the recognition rate on the degree of distortion, a number of experiments has been successfully conducted, confirming the above mentioned dependency of the two parameters, T_{60} and SMD. Further experiments have shown that ASR is barely affected by high-frequency reverberation, whereas low frequency reverberation has a detrimental effect on the recognition rate.

A literature survey concludes that, although several approaches exist which claim significant improvements, none of them fulfils the above mentioned practical implementation criteria. Within this thesis, a new approach entitled 'harmonicity-based feature analysis' (HFA) is proposed. It is based on three ideas that are derived in former chapters. Experimental results prove that HFA is able to enhance the recognition rate in reverberant environments. Even practical applicable results are achieved when HFA is combined with reverberant training. The method is further evaluated against three other approaches from the literature. Also combinations of methods are tested.

In a last chapter the two base technologies fundamental frequency (F_0) estimation and voiced unvoiced decision (VUD) are evaluated in reverberant environments, since they are necessary to run HFA. This evaluation aims to find one optimal method for each of these technologies. The results show that all F_0 estimation methods and also the VUD methods have a strong decreasing performance in reverberant environments. Nevertheless it is shown that HFA is able to deal with uncertainties of these base technologies as such that the recognition performance still improves.

Zusammenfassung

Bei der Überführung eines wissenschaftlichen Laborsystems zur automatischen Spracherkennung in eine reale Anwendung ergeben sich verschiedene praktische Problemstellungen, von denen eine der Verlust an Erkennungsleistung durch umgebende akustische Störungen ist. Im Gegensatz zu additiven Störungen wie Lüfterrauschen o. ä. hat die Wissenschaft bislang die Störung des Raumhalls bei der Spracherkennung nahezu ignoriert. Dabei besitzen, wie in der vorliegenden Dissertation deutlich gezeigt wird, bereits geringfügig hallende Räume einen stark störenden Einfluss auf die Leistungsfähigkeit von Spracherkennern.

Mit dem Ziel, die Erkennungsleistung wieder in einen praktisch benutzbaren Bereich zu bringen, nimmt sich die Arbeit dieser Problemstellung an und schlägt Lösungen vor. Der Hintergrund der wissenschaftlichen Aktivitäten ist die Erstellung von funktionsfähigen Sprachbenutzerinterfaces für Gerätesteuerungen im Wohn- und Büroumfeld, wie z. B. bei der Hausautomation. Aus diesem Grund werden praktische Randbedingungen wie die Restriktionen von embedded Computerplattformen in die Lösungsfindung einbezogen.

Die Argumentation beginnt bei der Beschreibung der raumakustischen Umgebung und der Ausbreitung von Schallfeldern in Räumen. Es wird theoretisch gezeigt, dass die Störung eines Sprachsignals durch Hall von zwei Parametern abhängig ist: der Sprecher-Mikrofon-Distanz (SMD) und der Nachhallzeit T_{60}. Um die Abhängigkeit der Erkennungsleistung vom Grad der Hallstörung zu ermitteln, wird eine Anzahl von Erkennungsexperimenten durchgeführt, die den Einfluss von T_{60} und SMD nachweisen. Weitere Experimente zeigen, dass die Spracherkennung kaum durch hochfrequente Hallanteile beeinträchtigt wird, wohl aber durch tieffrequente.

In einer Literaturrecherche wird ein Überblick über den Stand der Technik zu Maßnahmen gegeben, die den störenden Einfluss des Halls unterdrücken bzw. kompensieren können. Jedoch wird auch gezeigt, dass, obwohl bei einigen Maßnahmen von Verbesserungen berichtet wird, keiner der gefundenen Ansätze den o. a. praktischen Einsatzbedingungen genügt.

In dieser Arbeit wird die Methode Harmonicity-based Feature Analysis (HFA) vorgeschlagen. Sie basiert auf drei Ideen, die aus den Betrachtungen der vorangehenden Kapitel abgeleitet werden. Experimentelle Ergebnisse weisen die Verbesserung der Erkennungsleistung in halligen Umgebungen nach. Es werden sogar praktisch relevante Erkennungsraten erzielt, wenn die Methode mit verhalltem Training kombiniert wird. Die HFA wird gegen Ansätze aus der Literatur evaluiert, die ebenfalls praktischen Implementierungskriterien genügen. Auch Kombinationen der HFA und einigen dieser Ansätze werden getestet.

Im letzten Kapitel werden die beiden Basistechnologien Stimmhaft-Stimmlos-Entscheidung und Grundfrequenzdetektion umfangreich unter Hallbedingungen getestet, da sie Voraussetzung für die Funktionsfähigkeit der HFA sind. Als Ergebnis wird dargestellt, dass derzeit für beide Technologien kein Verfahren existiert, das unter Hallbedingungen robust arbeitet. Es kann allerdings gezeigt werden, dass die HFA trotz der Unsicherheiten der Verfahren arbeitet und signifikante Steigerungen der Erkennungsleistung erreicht.

Inhaltsverzeichnis

Abkürzungsverzeichnis

ACF	Auto Correlation Function
ACLag	Auto Correlation of Lagwindowing on Amplitude Spectrum
ACLS	Auto Correlation of Logarithmized Amplitude Spectrum
ACMWL	Auto Correlation through Multiple Window Length
ADU	Analog-Digital-Umsetzer
AEC	Acoustic Echo Cancellation
AM	Amplitudenmodulation
AMDF	Average Magnitude Difference Function
ASR	Automatic Speech Recognition
ASRU	IEEE Workshop on Automatic Speech Recognition and Understanding
BEA	Basic Extractor Algorithm
BSL	Blind Source Localization
BSS	Blind Source Separation
CCA	Complex Cepstrum Analysis
CFA	Conventional Feature Analysis
CHFA	Complex Harmonic Filter Analysis
CHIL	Computers in the Human Interaction Loop
CMN	Cepstral Mean Normalization
CMS	Cepstral Mean Subtraction
CMTF	Complex Modulation Transfer Function
DAU	Digital-Analog-Umsetzer
DCT	Discrete Cosine Transform
DFT	Diskrete Fourier Transformation
DIN	Deutsches Institut für Normung
DOA	Direction of Arrival
DP	Dynamische Programmierung
DR	Deletion Rate
DRR	Direct-to-Reverberation Ratio
DSB	Delay-and-Sum Beamformer
DSP	Digitaler Signalprozessor
DSV	Digitale Signalverarbeitung
DTFT	Discrete Time Fourier Transform
DTW	Dynamic Time Warping
ECESS	European Center of Excellence in Speech Synthesis
EGG	Elektroglottograph
ER	Error Rate
ETSI	European Telecommunications Standards Institute
EURASIP	The European Association for Signal Processing

EUSIPCO	European Signal Processing Conference
FCR	Fine Correct Rate
FER	Fine Error Rate
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FSB	Filter-and-Sum Beamformer
GCR	Gross Correct Rate
GER	Gross Error Rate
GMM	Gaussian Mixture Model
HERB	Harmonicity-based dEReverBeration
HFA	Harmonicity-based Feature Analysis
HMM	Hidden Markov Model
ICASSP	IEEE International Conference on Acoustics Speech and Signal Processing
ICSLP	International Conference on Spoken Language Processing
IEEE	Institute of Electrical and Electronics Engineers
IFFT	Inverse Fast Fourier Transform
IFHC	Instantanious Frequency of Harmonic Components
IIR	Infinite Impulse Response
IMTF	Inverse Modulation Transfer Function
IR	Insertion Rate
ISO	International Organization for Standardisation
ITU	International Telecommunications Union
IWAENC	International Workshop on Acoustic Echo and Noise Control
LP	Linear Prediction, lineare Prädiktion
LTFT	Long Time Fourier Transform
MCT	Multicondition Training
MFCC	Mel Frequency Cepstral Coefficients
MFB	Mel-Filterbank
MINT	Multiple-Input/Output INverse Theorem
MIPS	Mega Instructions per Second
MOS	Mean Opinion Score
MTF	Modulation Transfer Function
MSLP	Multi-step Forward Linear Prediction
NCCF	Normalized Cross Correlation Function
NR	Noise Reduction
O&A	Overlap-and-Add
OOV	Out-of-Vocabulary
PC	Personalcomputer
PCA	Principle Component Analysis
PDA	Pitch Detection Algorithm
PHIA	Periodicity/Harmonicity using Instantanious Amplitudes
PMC	Parallel Model Combination
RASTA	RelAtive SpecTrAl Processing
RASTI	Rapid Speech Transmission Index
REMOS	REverberation MOdeling for Speech Recognition
RejR	Rejection Rate

RIR	Room Impulse Response
RR	Recognition Rate
SDB	Superdirective Beamformer
SHS	Subharmonic Summation
SMD	Speaker-to-Microphone Distance
SMG	Stochastischer Markovgraph
SNR	Signal-to-Noise Ratio
SRM	Lautsprecher-Raum-Mikrofon
SRR	Signal-to-Reverberation Ratio
SIFT	Simplified Inverse Filter Tracking
STI	Speech Transmission Index
STPT	Short Term Power Spectrum Trajectory
TC-STAR	Technology and Corpora for Speech Translation
TME	Temporal Modulation Envelope
TPE	Temporal Power Envelope
TPEFA	Temporal Power Envelope Feature Analysis
UASR	Unified Approach for Speech Synthesis and Recognition
UE	Unvoiced Errors
UER	Unvoiced Errors Rate
UPC	Universitat Politècnica de Catalunya
VAD	Voice Activity Detector, Voice Activity Detection
VDA	Voicing Determination Algorithm
VE	Voiced Errors
VER	Voiced Errors Rate
VTS	Vector Taylor Series
VUD	Voiced Unvoiced Decision, Voiced Unvoiced Detector
WASPAA	IEEE Workshop on Applications of Signal Processing to Audio and Acoustics
WER	Word Error Rate
WRR	Word Recognition Rate

Benutzte Formelzeichen und Einheiten

Allgemein gilt: Signale und Systeme im Zeitbereich erscheinen in kleinen und im Frequenzbereich in großen Buchstaben. Im Zeitbereich besitzen sie die abhängigen Zeitvariablen t oder k, je nachdem ob das Signal kontinuierlich oder diskret ist. Im Frequenzbereich besitzen sie die abhängigen Frequenzvariablen ω, $e^{j\omega}$, n sowie den Frequenzindex n, je nachdem ob es sich um eine Fouriertransformierte, eine DTFT[1], eine DFT[2] (oder FFT[3]) oder um eine Fourierreihe handelt. Zusätzlich existiert für den Laplacebereich die abhängige Variable s und für den Bereich der $\mathcal{Z}$-Transformation die abhängige Variable z. Diese standardmäßigen abhängigen Variablen werden in der folgenden Liste der benutzten Formelzeichen nicht mit ausgeführt. Ändert sich die abhängige Variable innerhalb des Zeitbereichs von einem kontinuierlichen zu einem diskreten Signal bzw. umgekehrt, so ändert sich dar Buchstabe des Signals oder der Größe nicht. Gleiches gilt im Frequenzbereich. Bestehen andere Abhängigkeiten, erscheinen diese in Klammern (z. B. Abhängigkeit $p(h)$ des Luftdrucks p von der Höhe h).

Konventionen für ein Signal x

$x(t)$	kontinuierliches Signal im Zeitbereich
$x(k)$	diskretes Signal im Zeitbereich
$\tilde{x}$	Effektivwert von $x(t)$ oder $x(k)$
$\hat{x}$	Maximalwert von $x(t)$ oder $x(k)$
$\hat{x}$	Schätzwert von x
$\underline{X}_n$	Fourierkoeffizienten der Fourierreihe eines periodischen $x(t)$
$\underline{X}(\omega)$	Fouriertransformierte eines nichtperiodischen $x(t)$
$\underline{X}(n)$	DFT-Transformierte eines periodisch fortgesetzten $x(k)$
$\underline{X}(e^{j\omega})$	DTFT-Transformierte eines nichtperiodischen $x(k)$
$\underline{X}(z)$	$\mathcal{Z}$-Transformierte von $x(k)$
A	spektrale Amplitudendichte, Amplitudenspektrum
$x(a,k)$	Matrix von Frames von $x(k)$ für die Kurzzeit-Signalanalyse
$\boldsymbol{x}(a)$	Matrix von Frames von $x(k)$ in Vektor- bzw. Blockschreibweise
ψ_{xx}	Autokorrelationsfunktion einer Funktion x
ψ_{xy}	Kreuzkorrelationsfunktion der Funktionen x und y
S_{xx}	spektrale Leistungsdichte einer Funktion x
S_{xy}	spektrale Kreuzleistungsdichte der Funktionen x und y

[1] DTFT – Discrete Time Fourier Transform.
[2] DFT – Diskrete Fourier Transformation.
[3] FFT – Fast Fourier Transform.

Allgemeine Symbole

e	Euler'sche Zahl
π	Kreiskonstante
σ	Standardabweichung
σ^2	Varianz
μ	arithmetischer Mittelwert
ln	natürlicher Logarithmus ($\log_e$)
ld	dualer Logarithmus ($\log_2$)
lg	dekadischer Logarithmus ($\log_{10}$)
sin	Sinus
cos	Kosinus
rd	Rundungsoperator
$\mathcal{F}\{.\}$	Fouriertransformation
$\mathcal{H}\{.\}$	Hilberttransformation
$\mathcal{Z}\{.\}$	$\mathcal{Z}$-Transformation
$E\{.\}$	Erwartungswert
$\Re\{.\}$	Realteil
$\Im\{.\}$	Imaginärteil
$\lvert.\rvert$	Betrag

Akustische Ebene, Sprachsignale und Signalverarbeitung

a	Frameindex
a_u	stimmloser Frame
a_v	stimmhafter Frame
A	Anzahl der Frames
$A(p)$	Analysefunktion für die Grundfrequnzdetektion
A	äquivalente Absorptionsfläche
α	Absorptionsgrad
a_M	Gesamtübertragungsfaktor mehrerer linearer Teilssysteme von g_SRM
a_h	Proportionalitätsfaktor zwischen Schallenergiedichte und TPE[4]
b_{th}	Schwellwertparameter des VUD[5]
b_D	energiebezogener Faktor für die Beschreibung von h_D
b_G	energiebezogener Faktor für die Beschreibung von n_G
b_R	energiebezogener Faktor für die Beschreibung von h_R
c	Kanalindex einer Filterbank
C	Anzahl an Filterkanälen einer Filterbank
c	Ausbreitungsgeschwindigkeit von Schallwellen
c	Cepstrum, allgemein
c_s	Cepstrum eines ungestörten Sprachsignals s
c_x	Cepstrum eines gestörten Sprachsignals x
c_h	Cepstrum einer RIR[6] h

[4]TPE – Temporal Power Envelope.
[5]VUD – Voiced Unvoiced Detector.
[6]RIR – Room Impulse Response.

C_{50}	Sprachklarheitsmaß, Deutlichkeitsmaß
C_{80}	Musikklarheitsmaß, Durchsichtigkeitsmaß
C_{ASR}	raumakustisches Störmaß für die Spracherkennung
C	Kosten, allgemein
$C_{i,j}$	Kosten beim Übergang vom i-ten in den j-ten Zustand
δ	Dirac-Impuls
d	Dämpfungsfaktor
D	Dämpfungsmaß
D_{50}	Deutlichkeit
$e_{c,\mathrm{MFB}}$	Energie des c-ten MFB[7]-Kanals
$\vec{e}_{\mathrm{MFB}}$	Energievektor von Kanälen einer MFB
e_{Ph}	Anregungssignal eines Lautes, Zeitbereich
$\underline{E}_{\mathrm{Ph}}$	Anregungssignal eines Lautes, Frequenzbereich
e	zeitliche Einhüllende (TME[8]) eines Signals, allgemein
e_s	TME eines Sprachsignal s
$\mathrm{e}_{s,c}$	TME eines Sprachsignals s im Frequenzband des c-ten Filterkanals
e_x	TME eines gestörten Sprachsignals x
e_h	TME einer RIR h
e^2	zeitliche Einhüllende der Leistungsfunktion (TPE) eines Signals, allgemein
e_s^2	TPE eines Sprachsignals s
$\mathrm{e}_{s,c}^2$	TPE eines Sprachsignals s im Frequenzband des c-ten Filterkanals
e_x^2	TPE eines gestörten Sprachsignals x
e_h^2	TPE einer RIR h
$\mathrm{e}_{h,\mathrm{D}}^2$	TPE der Direktschallphase einer RIR h_{D}
$\mathrm{e}_{h,\mathrm{R}}^2$	TPE der Hallphase einer RIR h_{R}
$\mathrm{e}_{\mathrm{IMTF}}^2$	IMTF[9] im Zeitbereich, Impulsantwort der IMTF
e_x	LP-Fehlersignal eines gestörten Signals x
E_h	Energie der Harmonischen
E_{th}	Schwellwert der Energie der Harmonischen bei der VUD
η	Nutz-Stör-Verhältnis
f	Frequenz
Δf	Frequenzschritt, Frequenzauflösung eines diskreten Spektrums
f	Abtastrate, Sampling-Frequenz
f_M	Momentanwert einer veränderlichen Frequenz, Momentanfrequenz
f_m	Modulationsfrequenz
$f_{g,o}$	obere Grenzfrequenz
$f_{g,u}$	untere Grenzfrequenz
$f_{\mathrm{HP\ cut\ off}}$	Cut-Off-Frequenz eines Hochpasses zur Modifikation von RIRs
$f_{\mathrm{TP\ cut\ off}}$	Cut-Off-Frequenz eines Tiefpasses zur Modifikation von RIRs
F_0	Grundfrequenz
$F_{0,\max}$	Maximalwert der Grundfrequenz
$F_{0,\min}$	Minimalwert der Grundfrequenz

[7]MFB – Mel-Filterbank.
[8]TME – Temporal Modulation Envelope.
[9]IMTF – Inverse Modulation Transfer Function.

$F_{0,\mathrm{Est}}$	Schätzwert der Grundfrequenz
$F_{0,\mathrm{Ref}}$	Referenzwert der Grundfrequenz
Δ_{F_0}	Abweichung der Grundfrequenzschätzung
$\Delta_{F_0,\mathrm{FE}}$	Fine-Error-Limit der Abweichung der Grundfrequenzschätzung
$\Delta_{F_0,\mathrm{GE}}$	Gross-Error-Limit der Abweichung der Grundfrequenzschätzung
f_{d_u}	Beginn (Frequenz) des Übergangsbereichs bei der spektralen Synthese für stimmlose Frames
f_{w_u}	Breite (Frequenz) des Übergangsbereichs bei der spektralen Synthese für stimmlose Frames
f_{d_v}	Beginn (Frequenz) des Übergangsbereichs bei der spektralen Synthese für stimmhafte Frames
f_{w_v}	Breite (Frequenz) des Übergangsbereichs bei der spektralen Synthese für stimmhafte Frames
FI	Fortsatzintervall
g	Impulsantwort eines Systems, allgemein
g_{S}	Impulsantwort eines Systems, das ein Quellsignal s_{S} in eine Luftdruckschwankung $p_{s,\mathrm{Q}}$ umwandelt
g_{LS}	Impulsantwort des Systems eines Lautsprechers, der ein verstärktes Quellsignal s_{LS} in eine Luftdruckschwankung $p_{s,\mathrm{Q}}$ umwandelt
g_{M}	Impulsantwort des Systems eines Mikrofons, das eine Luftdruckschwankung p_{M} in ein Signal x_{M} umwandelt
g_{DAU}	Impulsantwort des Systems eines DAU[10]
g_{ADU}	Impulsantwort des Systems eines ADU[11]
$g_{\mathrm{Amp,LS}}$	Impulsantwort des Systems eines Lautsprecherverstärkers
$g_{\mathrm{Amp,M}}$	Impulsantwort des Systems eines Mikrofonverstärkers
g_{Raum}	Impulsantwort des akustischen Systems Raum zwischen Quelle und Empfänger
g_{SRM}	Impulsantwort des SRM-Systems
γ	Exponent bei der spektralen Synthese
γ	Bündelungsgrad bzw. Bündelungsmaß in dB
$\Gamma(\theta,\phi)$	Richtungsfunktion
h	Raumimpulsantwort, Impulsantwort des SRM-Systems
h_m	Raumimpulsantwort zwischen Quelle und m-ten Mikrofon
$\underline{H}$	Übertragungsfunktion des SRM-Systems
$\underline{H}_m$	Übertragungsfunktion zwischen Quelle und m-ten Mikrofon
h_{D}	Impulsantwort des Teilsystems von h, das den Direktschall beschreibt
h_{R}	Impulsantwort des Teilsystems von h, das den Hall beschreibt
$h_{\mathrm{R},i}$	Impulsantwort des Teilsystems von h, das die i-te diskrete Reflexion beschreibt
h_{mod}	modifizierte RIR
H	Hallabstand
H_w	wirksamer Hallabstand
I	Schallintensität

[10] DAU – Analog-Digital-Umsetzer.
[11] ADU – Digital-Analog-Umsetzer.

i, j, l	Index, allgemein
k	Index der diskreten Zeit
K	Anzahl von Abtastpunkten eines Signalabschnittes
κ	Adiabatenexponent
κ	Index der Verschiebung der diskreten Zeit
L	Pegel, allgemein
L_p	Schalldruckpegel
L_{w_D}	Pegel der Schallenergiedichte des Direktschallfeldes
L_{w_R}	Pegel der Schallenergiedichte des Hallfeldes
L_h	Pegel der Leistungsfunktion der RIR h^2
L_Schr	Pegel des Schröderintegrals
$L_{b_{th}}$	Schwellwertparameter des VUD als Pegelangabe
$L_{E_{th}}$	Schwellwert der Energie der Harmonischen beim VUD in Pegelschreibweise
L	Interpolationsfaktor (Upsampling Factor), allgemein
L_c	Interpolationsfaktor im Frequenzbereich zwischen Filterbankkanälen c
m	Index des m-ten Mikrofons
M	Anzahl an Mikrofonen
$\underline{M}_s(\omega_m)$	Modulationsspektrum eines Sprachsignals s
$\underline{M}_x(\omega_m)$	Modulationsspektrum eines gestörten Sprachsignals x
$\underline{M}_h(\omega_m)$	Modulationsspektrum einer RIR h
$\underline{M}_{h,\mathrm{R}}(\omega_m)$	Modulationsspektrum der Hallphase einer RIR h_R
$\underline{M}_{s,c}(\omega_m)$	Modulationsspektrum eines Sprachsignals s im Frequenzband des Kanals c
$\underline{m}(\omega_m)$	komplexe Modulationsübertragungsfunktion (CMTF[12])
$m(\omega_m)$	Modulationsübertragungsfunktion (MTF[13])
$m^{-1}(\omega_m)$	inverse Modulationsübertragungsfunktion (IMTF)
M	Dezimationsfaktor (Downsampling Factor), allgemein
M_k	Dezimationsfaktor im Zeitbereich (diskrete Zeit k)
N	Anzahl, allgemein
n	Index der diskreten Frequenz
N	FFT- bzw. DFT-Breite
$n_{h,i}$	i-ter harmonischer Frequenzindex
n	additives Störgeräusch, Zeitbereich
$\underline{N}$	additives Störgeräusch, Frequenzbereich
n_G	Gaußrauschen
n_h	modelliertes Rauschen der Hallphase einer RIR
N_ges	Gesamtanzahl an Frames
$N_{u,\mathrm{Ref}}$	Anzahl stimmloser Frames
$N_{v,\mathrm{Ref}}$	Anzahl stimmhafter Frames
N_VE	Anzahl an Voiced Errors
N_UE	Anzahl an Unvoiced Errors
ω	Kreisfrequenz
ω_m	Modulationskreisfrequenz
ω_M	Momentankreisfrequenz

[12] CMTF – Complex Modulation Transfer Function.
[13] MTF – Modulation Transfer Function.

p	Druck
p_-	statischer Druck
$p_\sim$	Druckschwankung, Schalldruck
p_D	Schalldruckkomponente des Direktschallfeldes
p_R	Schalldruckkomponente des Hallfeldes
p_0	Schalldruck an der genormten Hörschwelle
$p_{s,\mathrm{Q}}$	Schalldruck-Zeit-Funktion des Sprachsignals s an der Quelle
$p_{s,\mathrm{M}}$	Schalldruck-Zeit-Funktion des Sprachsignals s am Mikrofon
$p_{n,\mathrm{M}}$	Schalldruck-Zeit-Funktion des Geräuschsignals n am Mikrofon
p_M	Schalldruck-Zeit-Funktion am Mikrofon
P_Q	Schallleistung einer Quelle
q	Quefrenz
Q	Volumengeschwindigkeit
Q	Schalleindruck
r	Abstand, Sprecher-Mikrofon-Abstand (SMD[14])
r_Test	SMD zur Erkennungsphase
r_Train	SMD zur Trainingsphase
r_R	Hallradius
R	Hallmaß
R_w	Raumeindrucksmaß
R_s	stoffspezifische Gaskonstante
ρ	Reflexionsgrad
ϱ	Dichte
ϱ_-	statische Dichte
$\varrho_\sim$	Dichteschwankung
s	Sprachsignal, Zeitbereich
s_S	Sprachsignal, ausgesendet von einem Sprecher
s_LS	Sprachsignal, ausgesendet von einem Lautsprecher
$\underline{S}$	Sprachsignal, Frequenzbereich
s'	verhalltes Sprachsignal
s'_D	Direktschallkomponente von s'
s'_R	Diffusschall-, Hallkomponente von s'
$s'_{\mathrm{R},i}$	Hallkomponente einer diskreten Reflexion von s'
s_Ph	Sprachsignal eines Lautes, Zeitbereich
$\underline{S}_\mathrm{Ph}$	Sprachsignal eines Lautes, Frequenzbereich
s_h	harmonische Komponenten eines Sprachsignals
s_n	nichtharmonische Komponenten eines Sprachsignals
σ^2_G	Leistung eines mittelwertfreien Gaußrauschens
t	Zeit
τ	Zeit, Verschiebung in Zeitrichtung
T	Zeitintervall, -abschnitt
T	Periodendauer
$t_\mathrm{cut\ off}$	Cut-Off-Zeit zur Modifikation von RIRs
T_{FI}	Zeitdauer eines Framintervalls

[14] SMD – Speaker-to-Microphone Distance.

T_K	Zeitdauer eines Frames
T_N	Zeitdauer des DFT- bzw. FFT-Eingangsfensters
T_0	Grundperiodendauer
$T_{0,\max}$	Maximalwert der Grundperiodendauer
$T_{0,\min}$	Minimalwert der Grundperiodendauer
T_{60}	Nachhallzeit
$T_{60,\mathrm{Test}}$	Nachhallzeit zur Erkennungsphase
$T_{60,\mathrm{Train}}$	Nachhallzeit zur Trainingsphase
T_{20}, T_{30}	Nachhallzeit bei Meßgeräten, gemessen über 20 bzw. 30 dB
ϑ	Temperatur in °C
T	absolute Temperatur
T_{-}	statische absolute Temperatur
$T_{\sim}$	Schwankung der absoluten Temperatur
V	Volumen
$\vec{v}$, v	Schallschnelle
v_{Ph}	Impulsantwort des Vokaltrakts eines Lautes
$\underline{V}_{\mathrm{Ph}}$	Übertragungsfunktion des Vokaltrakts eines Lautes
VUD_{Est}	Schätzwert des VUD
VUD_{Ref}	Referenzwert des VUD
w	Schallenergiedichte
w_{D}	Schallenergiedichte des Direktschallfeldes
w_{R}	Schallenergiedichte des Hallfeldes
$w_{h_{\mathrm{R}}}$	momentane, mittlere Schallenergiedichte der Hallphase einer RIR h
W_h	Energie der RIR h
$W_{h_{\mathrm{D}}}$	Energie des Teilsystems der RIR h, das den Direktschall beschreibt
$W_{h_{\mathrm{R}}}$	Energie des Teilsystems der RIR h, das den Hall beschreibt
$W_{h_{\mathrm{nutz},t_k}}$	Energie der nützlichen Reflexionen der RIR h bis zu einer kritischen Verzögerungszeit t_k
$W_{h_{\mathrm{stör},t_k}}$	Energie der störenden Reflexionen der RIR h ab einer kritischen Verzögerungszeit t_k
$W_{h_{\mathrm{R},i}}$	Energie der diskreten Reflexionen in der RIR h
W_{e_h}	Energie des TPEs e_h^2 der RIR h
w_{inv}	Impulsantwort eines inversen Systems zu h
$w_{m,\mathrm{inv}}$	Impulsantwort eines inversen Systems zu h_m
$\underline{W}_{\mathrm{inv}}$	Übertragungsfunktion eines inversen Systems zu $\underline{H}$
$\underline{W}_{m,\mathrm{inv}}$	Übertragungsfunktion eines inversen Systems zu $\underline{H}_m$
w_{Hann}	Von-Hann-Fenster
$W_{u,L}$	Ausblendfunktion der spektralen Synthese für stimmlose Frames
$W_{u,H}$	Einblendfunktion der spektralen Synthese für stimmlose Frames
$W_{v,L}$	Ausblendfunktion der spektralen Synthese für stimmhafte Frames
$W_{v,H}$	Einblendfunktion der spektralen Synthese für stimmhafte Frames
$\mathbf{W}$	Transformationsmatrix
w	Kostenfuntion beim Viterbi-VUD, allgemein
w_{VU}	Kosten beim Viterbi-VUD beim Übergang von stimmhaften zu stimmlosen Zuständen

w_{UV}	Kosten beim Viterbi-VUD beim Übergang von stimmlosen zu stimmhaften Zuständen
w_{UU}	Kosten beim Viterbi-VUD beim Übergang von stimmlosen zu stimmlosen Zuständen
x	verhalltes Sprachsignal ($x = s$' unter der Annahme $n(t) = 0$)
x	Mikrofonsignal, Eingangssignal für die DSV[15], Zeitbereich
$\underline{X}$	Mikrofonsignal, Eingangssignal für die DSV, Frequenzbereich
x_m	Mikrofonsignal des m-ten Mikrofons
x_{M}	Mikrofonsignal vor der Verstärkung des Mikrofonverstärkers
$x(a,k)$, $\boldsymbol{x}(a)$	Matrix von Zeitsignalframes eines Eingangssignals
x_{D}	Direktschallkomponente von x
x_{R}	Hallkomponente von x
$x_{\mathrm{D},h}$	harmonische Anteile der Direktschallkomponente von x
$x_{\mathrm{D},n}$	nichtharmonische Anteile der Direktschallkomponente von x
$x_{\mathrm{R},h}$	harmonische Anteile der Hallkomponente von x
$x_{\mathrm{R},n}$	nichtharmonische Anteile der Hallkomponente von x
x_n	nichtharmonische Anteile von x
$x_{h_{\mathrm{mod}}}$	verhalltes Sprachsignal durch eine modifizierte RIR h_{mod}
x_{DSB}	Ausgangssignal eines Dealy-and-Sum-Beamformers
$\underline{x}_{a,c}$	Analytisches Signal eines Frequenzbandes des c-ten Filterkanals
X_S	synthetisiertes Spektrum
$X_{S,h}$	synthetisiertes harmonisches Spektrum
$X_{S,v}$	synthetisiertes Spektrum für stimmhafte Frames
$X_{S,u}$	synthetisiertes Spektrum für stimmlose Frames
X_h	Spektrum von Harmonischen
$X_{\mathrm{e}_x^2}$	TPE-Spektrum nach TPEFA[16]
$X_{\mathrm{e}_S^2}$	TPE-Spektrum nach HFA[17]+TPEFA
$X_{\hat{\mathrm{e}}_s^2}$	restauriertes TPE-Spektrum nach TPE-Enthallung durch IMTF
$\vec{\boldsymbol{x}}(a)$	Merkmalvektor, extrahiert aus dem Frame a
$\vec{\boldsymbol{x}}'(a)$	Merkmalvektor mit dynamischen Merkmalen (Frame a)
$\vec{\boldsymbol{x}}''(a)$	Merkmalvektor nach Merkmaltransformation (Frame a)
$\vec{\boldsymbol{x}}_{\mathrm{MFB}}$	MFB-Merkmalvektor
$x_{c,\mathrm{MFB}}$	Vektorkomponente des c-ten MFB-Kanals von $\vec{\boldsymbol{x}}_{\mathrm{MFB}}$
$\vec{\boldsymbol{x}}_{\mathrm{MFCC}}$	MFCC[18]-Merkmalvektor
$x_{q,\mathrm{MFCC}}$	Vektorkomponente der q-ten Quefrenz von $\vec{\boldsymbol{x}}_{\mathrm{MFCC}}$
Z	Mel-Skala
Z	Schallimpedanz

[15] DSV – digitale Signalverarbeitung
[16] TPEFA – Temporal Power Envelope Feature Analysis.
[17] HFA – Harmonicity-based Feature Analysis.
[18] MFCC – Mel Frequency Cepstral Coefficients.

Klassifikation und Evaluation

C	Anzahl an korrekten Erkennungen
D	Anzahl an Auslassungen (Deletions)
I	Anzahl an Einfügungen (Insertions)
K	Anzahl an Klassen eines Erkenners
m	Anzahl an Gaußverteilungen in einem GMM[19]
N_O	Anzahl an OOV[20]-Testwörtern
R	Menge aller Erkennungsergebnisse (Resultat)
R_k	k-te erkannte Klasse
S	Anzahl an Verwechslungen (Substitutions)
T	Menge aller Eingabeklassen (Testklassen)
T_k	k-te Eingabeklasse (Testklasse)
T_O	OOV-Testklasse
W	Wortvokabular
W_k	k-te Klasse aus dem Wortvokabular W

Einheiten

°C	Grad Celsius
dB	Dezibel
K	Kelvin
kg	Kilogramm
ms, s	Millisekunden, Sekunden
mm, cm, m	Millimeter, Zentimeter, Meter
m^3	Kubikmeter
N	Newton
Pa	Pascal

[19] GMM – Gaussian Mixture Model.
[20] OOV – Out-of-Vocabulary.

1 Einführung

1.1 Forschungsgegenstand

Sprache, die zwischen Menschen die gängigste Kommunikationsform darstellt, wird in Literatur und Film immer wieder als technisch beherrschbare Modalität der Mensch-Maschine-Schnittstelle beschrieben. In der Realität der heutigen Zeit sind derartige Systeme im Verhältnis zu ihren potentiellen Anwendungsmöglichkeiten jedoch eher seltene Ausnahmen. Beispiele sind Diktiersysteme, Sprachinterfaces beim Mobiltelefon oder bei Komfortelektronik im Automobil, Sprachdialogsysteme bei Telefonhotlines sowie auch die Steuerung eines Operationsmikroskopes durch den Operateur. Die Akzeptanz derartiger Anwendungen ist mehr oder weniger groß. Das liegt u. a. daran, dass die Qualität der Spracherkennung den Erwartungen der Nutzer, die ja als Gegenüber einen in vielen akustischen Szenarien gut verstehenden Menschen gewohnt sind, nicht ausreichend gerecht wird. Im Jahre 1999 wurde durch die Zeitschrift Computer-Bild ein professioneller Test von zehn Diktiersystemen beauftragt, der das ernüchternde Ergebnis lieferte, dass aus Sicht des Anwenders fünf Systeme die Note ausreichend (4) und fünf die Note mangelhaft (5) erhalten[1]. Die genannten Anwendungen sind dadurch gekennzeichnet, dass sie Funktionen steuern, die keine kritischen Bereiche betreffen, sondern sie wirken unterstützend, z. B. an Stellen, wo beide Hände anderweitig benötigt werden. Fehlerkennungen dürfen keine sicherheitsrelevanten Auswirkungen haben und der Nutzer muss ihr Auftreten akzeptieren können. In Anwendungen im Wohn- und Büroumfeld ist dem Autor derzeit kein erfolgreich am Markt positioniertes, kommerziell verfügbares System bekannt, welches mit einer Sprachsteuerung ausgerüstet ist[2]. Dabei besteht gerade hier ein hohes Potential an Anwendungen, bei denen durch eine Sprachsteuerung eine Steigerung des Komforts erzielt werden könnte. Aber nicht nur Komfort, sondern auch die Unterstützung von behinderten Menschen bietet ein breites Anwendungsfeld im Haushalt und weiteren Umgebungen. Beispiele wären die Steuerung von Licht, Jalousie, Türöffner, Unterhaltungselektronik etc. per Sprache. Im Jahre 2003 wurde von einem renommierten Hersteller für Haushaltselektronik die Entwicklung einer Dunstabzugshaube geplant, die eine Sprachbedienung mit

[1] Ein neuerer Test von Diktiersystemen ist dem Autor nicht bekannt. In der etwas jüngeren Arbeit [Wer08] findet man jedoch die Bemerkung, dass sich trotz technischem Fortschritt an dieser Situation bislang nur wenig geändert hat.

[2] Der Leser kann sich einen Einblick bei einschlägigen Märkten für Haushalts- und Unterhaltungselektronik verschaffen.

einem Satz an Sprachkommandophrasen für integrierte Funktionen, wie die Steuerung von Licht oder Ventilator, vorsah. Um einen geeigneten Hersteller für Spracherkenner zu finden, wurde eine Evaluation beauftragt, bei der zehn kommerziell verfügbare Spracherkenner umfangreich unter realistischen akustischen Küchenszenarien getestet wurden [MHK+03]. Das Ergebnis dieser Evaluation war, dass keiner der Testteilnehmer den vorgegebenen Qualitätsansprüchen gerecht wurde, sodass die Entwicklung des Gerätes mit Sprachsteuerung nicht fortgeführt wurde. Bemerkenswert ist dabei, dass bei diesem Test die zu erreichende, vom Nutzer akzeptierte Erkennungsrate mit 85 %[3] nicht etwa besonders hoch festgelegt wurde. Kommerzielle Erkenner werben mit Erkennungsraten von $> (95 \ldots 98)$ %[4]. In wissenschaftlichen Veröffentlichungen wird von ähnlichen Erkennungsraten berichtet. [Dro08] stellt z. B. drei Erkenner gegenüber, die für eine bestimmte Aufgabe unter Laborbedingungen jeweils $(99,62 \pm 0,02)$ % Erkennungsrate erreichen, die allerdings bei Hinzufügen von Geräusch rapide einbricht. Als Grund für das Versagen der Systeme wird in [MHK+03] besonders der Informationsverlust des Sprachsignals durch akustische Störungen auf dem Weg vom Sprecher zum Mikrofon verantwortlich gemacht. Vergleicht man die oben genannten verfügbaren Applikationen mit der Anwendung im Haushalt, stellt man zunächst fest, dass der Abstand zwischen Sprecher und Mikrofon gering und meist optimal ist. Dies ist durch die Benutzung von Headsets, Telefonhörern oder eine kopfnahe Montage des Mikrofons (z. B. A-Säule im PKW) gegeben. Zusätzlich erleichtert ein dezidiertes, begrenzt andauerndes Einschalten des Mikrofons bei bestimmten Dialogpositionen das Verhindern von sogenannten Insertions[5]. Beispiele dafür sind die Push-to-Talk-Taste im Kfz oder beim Mobiltelefon bzw. die Menüstruktur eines Dialogs bei einer Telefonhotline. Beides ist im Haushaltsbereich kaum umzusetzen; der Nutzer würde weder die Benutzung eines Headsets akzeptieren noch ist die Verwendung einer Aktivierungstaste sinnvoll, da in dem Fall die Sprachsteuerung obsolet wäre[6]. Durch den größeren Sprecher-Mikrofon-Abstand (engl.: Speaker-to-Microphone Distance – SMD) sinkt die Leistung des Sprachsignals (Nutzsignal) im Vergleich zu einer gleichmäßig im Raum verteilten Störung, der Signal-Rausch-Abstand (engl.: Signal-to-Noise Ratio – SNR) verringert sich. Schlussfolgernd darf angenommen werden, dass derzeitige Systeme nur bei Einhaltung eines geringen SMD $< (30 \ldots 50)$ cm einsetzbar sind. Bei größeren SMDs versagen sie aufgrund des steigenden SNRs.

Die vorliegende Arbeit nimmt sich dieser Problemstellung an. Dabei ist das Verständnis des akustischen Szenarios auf der Sprecher-Mikrofon-Strecke von Bedeutung, das in Abbildung 1.1 schematisch dargestellt ist. Der direkte Schall des Sprachsignals $s(t)$ ist

[3] Dieser Wert wurde zuvor durch Usability-Experimente mit Versuchspersonen bestimmt [MHK+03]. Er wird für die vorliegende Arbeit als Richtwert betrachtet.

[4] Wäre nicht allgemein bekannt, dass Spracherkennung immer fehlerbehaftet ist, würde man vermutlich auch 100 % angeben.

[5] Mit Insertions werden Reaktionen des Systems auf akustische Ereignisse bezeichnet, die nicht zum Vokabular der Anwendung gehören (vgl. Abschnitte 2.3.1 und 2.3.4).

[6] Bestenfalls könnten Anwesenheitssensoren zur Aktivierung die Fehleranfälligkeit mindern.

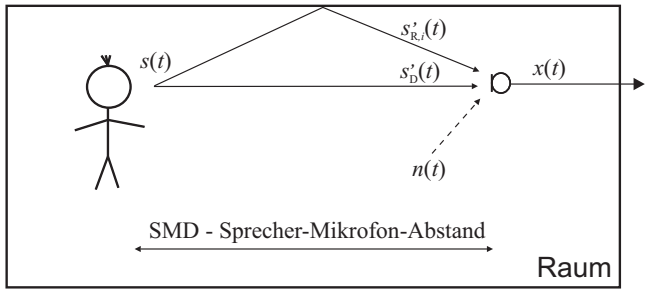

Abbildung 1.1 – Schema des akustischen Szenarios für die Spracherkennung in einem Raum. Die Störungen Raumhall und Geräusch werden angedeutet.

als Signalpfeil angedeutet und wird am Mikrofon als $s'_D(t)$ empfangen. Der Raumhall wird durch einen an einer Wand reflektierten Signalpfeil angedeutet, der symbolisch für multipel verteilte Reflexionen im Raum steht. Er wird mit $s'_R(t)$ bezeichnet (die dargestellte diskrete Reflexion hat die Bezeichnung $s'_{R,i}(t)$). Der gestrichelte Signalpfeil deutet das Umgebungsrauschen $n(t)$ an, das im Raum konstant verteilt angenommen wird. In der Abbildung erkennt man, dass sich zum Nutzsignal $s'_D(t)$ noch die (vermutlich) störend wirkenden Signale $s'_R(t)$ und $n(t)$ addieren (vgl. auch Abbildungen 3.7 und 3.8)

$$x(t) = s'_D(t) + s'_R(t) + n(t). \tag{1.1}$$

Historisch hat sich die Wissenschaft zunächst der Behandlung der Geräuschstörung $n(t)$ gewidmet, die offensichtlich zur Verschlechterung der Signalqualität führt. Aus diesem Grund liegen erste wissenschaftliche Arbeiten auf diesem Bereich bereits weit zurück in den 1970/80er Jahren (z. B. [Bol79]) und reichen somit fast bis an die Anfänge der digitalen Sprachsignalverabeitung (vgl. Abschnitt 1.3). Seitdem gibt es dazu eine Fülle von weiteren Arbeiten, sodass von einer wissenschaftlichen Sättigung der Thematik Robustheit gegen Störgeräusche ausgegangen werden kann. Nicht ganz so offensichtlich scheint der Einfluss des Halls zu sein, obwohl raumakustische Effekte seit langem nahezu vollständig untersucht sind (z. B. [Kut00]). Für die Sprachsignalverarbeitung ist die Behandlung raumakustischer Effekte allerdings noch ein junges Thema. Natürlich wird ebenfalls zeitig der Einfluss des Raumes gelegentlich verbal erwähnt. Genaue Untersuchungen sowie erste Ansätze zur Robustheitssteigerung gegen Raumhall oder Ansätze zur Enthallung reichen jedoch kaum mehr als fünf bis zehn Jahre zurück. Die Problematik darf, wie später dargestellt wird, als nahezu ungelöst angesehen werden. Deshalb widmet sich diese Arbeit der Untersuchung des Verhaltens von Kommandoworterkennern unter raumakustischen Umgebungsbedingungen sowie der Entwicklung von robustheitssteigernden Maßnahmen gegen Raumhall.

1.2 Randbedingungen: Anforderungen an Kommandoworterkenner

Bei der Entwicklung einer robustheitssteigernden Maßnahme gegen Raumhall sollen von vornherein praxisbezogene Randbedingungen beachtet werden. Raumhall ist erwartungsgemäß nicht das einzige Problem, mit dem man sich auseinandersetzen muss, wenn ein Laborsystem eines Kommandoworterkenners in eine reale Anwendung für Gerätesteuerungen überführt und erfolgreich eingesetzt werden soll. In die Überlegungen für einen neuen Ansatz müssen alle relevanten Randbedingungen einbezogen werden; es nützt bspw. wenig, wenn ein Enthallungsverfahren Signale zwar besonders gut enthallen kann, dazu aber Adaptionszeiten von einer Stunde Sprachmaterial benötigt (so z. B. der Ansatz HERB [NKM07]). Es ist deshalb nötig zu analysieren, welche Anforderungen an derartige Systeme gestellt werden. Dabei sind zunächst nicht technische Gegebenheiten relevant, sondern vor allem Anforderungen des Kunden, der ein Gerät nur dann kauft, wenn er mit Preis und Leistung zufrieden ist. Die technischen Randbedingungen werden davon abgeleitet. Die wichtigsten Anforderungen für reale Anwendungen sind wie folgt beschrieben.

Funktionale Anforderungen Für das Design eines Gerätes sowie seiner Sprachbedienungseinheit entstehen funktionale Anforderungen, die zu Randbedingungen der verwendeten Sprachtechnologien führen:

- **einfacher, intuitiver Dialog** – Der Nutzer sollte das Gerät möglichst intuitiv bedienen können, ohne es zu kennen und ohne die Bedienungsanleitung zu lesen. Die dazu nötige Dialogstruktur sollte zusätzlich überschaubar und flachhierarchisch sein. Dadurch kommt der Nutzer schnell zum Ziel (Task-Completion-Time).
- **akzeptable Erkennungsleistung** – Es ist nicht erforderlich, 100 % korrekt zu erkennen. In [MHK+03] wurde, wie bereits erwähnt, eine akzeptable Erkennungsrate von > 85 % festgelegt. Für diese Arbeit soll ein Wert von 90 % gelten. Der Nutzer akzeptiert verschiedene Erkennungsfehler in gewissen Grenzen (genauere Erläuterungen in Abschnitt 2.3.1). Dabei kann folgende Wichtung festgestellt werden:
  (i) Rückweisung eines Kommandos: Sowohl kooperative als auch weniger kooperative Nutzer akzeptieren in gewissen Grenzen, wenn ein Kommando wiederholt werden muss, da dies in der zwischenmenschlichen Kommunikation auch stattfindet.
  (ii) Verwechslung eines Kommandos: Wird normalerweise nicht akzeptiert, allerdings kann der Nutzer die fehlerhaft ausgelöste Aktion abbrechen, da er gerade beim Bedienen des Gerätes ist.
  (iii) Insertion (Reaktion auf Schallereignisse, die nicht zum Vokabular gehören): Wird keinesfalls akzeptiert, da das Gerät bei fremden Schallereignissen eine Aktion auslöst, ohne dass es bedient wird. Der Nutzer kann u. U. nicht eingreifen, wenn er bspw. nicht vor Ort ist.

- **integrierte Lösung** – Das Sprachinterface muss in das Gerät integriert werden können (eingebettete Lösung, engl.: embedded). Aus funktionalen Gesichtspunkten ist auch eine zentrale Serverlösung denkbar, allerdings muss ein zusätzlicher Kommunikationskanal zwischen Mikrofon und Server sowie zwischen Server und Gerätesteuerung zur Verfügung gestellt werden.
- **schnelle Reaktionszeit** – Die Zeit, die der Erkenner zur Erkennung eines Befehls benötigt, sollte < 1 s sein. Dieser Wert wird für die vorliegende Arbeit als Obergrenze festgelegt, bei der noch von einer Reaktion in Echtzeit gesprochen werden kann.

Kostenanforderungen Der Nutzer ist bereit, für die zusätzliche Funktionalität der Sprachbedienung einen höheren Preis zu zahlen. Allerdings ist die Höhe des Mehrpreises begrenzt[7]. Will man eine ganze Kücheneinrichtung zentral steuern, kommt preislich die Installation einer Serverlösung in Frage. Bei einer Sprachbedienung für ein einzelnes Gerät ist nur die Benutzung eines embedded Prozessors sinnvoll, der aus Preisgründen derzeit nur auf Festkommatechnologie basieren darf (z. B. Festkomma-DSP[8]). Für preisgünstige Geräte (z. B. $<$ €100,–) kommen nur noch Mikrocontroller in Frage, deren Hardwareressourcen wesentlich geringer sind[9]. Für diese Arbeit wird ein State-of-the-Art-Festkomma-DSP als preislich akzeptable Lösung angesehen. Die Lösung soll möglichst mit nur einem Mikrofon auskommen. Mehrere Mikrofone verursachen größere Kosten durch den Preis des Mikrofons sowie zusätzliche ADU[10]-Hardware.

Implementierungsbedingungen Der oben angesprochene State-of-the-Art-Festkomma-DSP markiert im Vergleich zu einem Laborsystem verschiedene Randbedingungen durch seine begrenzten Ressourcen. Dies betrifft die vier Bereiche:

- **Rechenleistung** – Heutige DSPs besitzen bereits bemerkenswerte Rechengeschwindigkeiten. So sind im preislich vertretbaren Bereich schon mehrere hundert MIPS[11] möglich. Dennoch liegt der DSP deutlich unter der Leistungsfähigkeit eines derzeit verfügbaren PCs[12] oder Servers. Bei der Portierung des Laborsystems auf einen DSP muss darauf geachtet werden, dass mit diesen begrenzten Ressourcen die Reaktionszeit eingehalten wird. Ggf. muss die Implementierung optimiert werden, was

[7] Ein Sprachbedienungsmodul kann den Verkaufspreis einer Dunstabzugshaube von bspw. €400,– auf €500,– steigern, bei teureren Produkten kann die absolute Preissteigerung noch etwas erhöht werden. Weitere Preissteigerungen sind kaum möglich.

[8] DSP – Digitaler Signalprozessor.

[9] Diese Angaben können sich bei Weiterentwicklung der Technik ändern. Es ist anzunehmen, dass in Zukunft wesentlich leistungsfähigere embedded Prozessoren zu günstigen Preisen verfügbar sein werden.

[10] ADU – Analog-Digital-Umsetzer.

[11] MIPS – Mega Instructions per Second.

[12] PC – Personalcomputer.

zum Verlust von Erkennungsleistung führen kann.

- **Rechengenauigkeit** – Für einige Algorithmen in Spracherkennungssystemen, wie bspw. die Berechnung von Wahrscheinlichkeiten bei HMMs, sind teilweise hohe Rechengenauigkeiten erforderlich, die durch die Verwendung von Fließkommaarithmetik gegeben sind. Bei einem Festkomma-DSP stehen verschiedene Werkzeuge zur Verfügung, um die Fließkommagenauigkeit zu erreichen. Als erstes ist es möglich, die Fließkommaoperationen durch Softwareroutinen zu emulieren. Dies stellt die gängigste Methode dar, allerdings geht sie auf Kosten der Rechenleistung (z. B. Faktor 100 pro Fließkommaoperation), die, wie erwähnt, im Vergleich zu PCs bereits reduziert ist. Bei einer zu großen Anzahl von Fließkommaoperationen bieten sich weitere Strategien, wie das Skalieren oder Tabellieren, an. Diese sind aber ggf. mit einem Verlust an Rechengenauigkeit verbunden, sodass u. U. die Erkennungsleistung sinkt.
- **Datenspeicher** – ein On-Chip-Datenspeicher ist eine erhebliche Ressourceneinschränkung gegenüber dem Laborsystem. Es darf mit einigen zehn Kilobyte gerechnet werden, wohingegen PC-Software im Megabyte Bereich arbeitet. Bei der Portierung können zwei Strategien verfolgt werden: die Optimierung des Speicherverbrauchs der Software, was ggf. wieder mit einem Verlust an Erkennungsleistung verbunden ist, sowie die Verwendung eines Off-Chip-Datenspeichers, der die Kosten wieder leicht erhöht und meist eine größere Zugriffszeit besitzt.
- **Programmspeicher** – hier gilt das gleiche wie beim Datenspeicher, allerdings ist aus implementierungstechnischen Gründen anzustreben, ohne Off-Chip-Programmspeicher auszukommen.

Frühere Arbeiten des Autors befassen sich mit der Portierung von Laborerkennern auf embedded Plattformen [HBK+02, PHHJ03, PHHJ04]. Für diese Arbeit liegt der Fokus auf der Robustheit der Erkennungsgenauigkeit. In Hinblick auf die praktische Einsetzbarkeit robustheitssteigernder Verfahren ist es dennoch nötig, diese ressourcenbegrenzenden Randbedingungen bei der Auswahl oder der Entwicklung von Ansätzen zu beachten. Es ist eine möglichst ressourcensparende Methode zu finden, wie sie z. B. in Kapitel 6 vorgestellt wird.

Reale akustische Umgebungsbedingungen Im vorigen Abschnitt wurde bereits ein reales akustisches Szenario schematisch dargestellt. Im Vergleich zu definierten Laborbedingungen sind in der Realität Störungen wie Rauschen und Hall vorzufinden, wobei aus Gründen der aktuellen wissenschaftlichen Relevanz, wie erwähnt, der Fokus in dieser Arbeit auf der Behandlung von Raumhall liegt. Für diese Arbeit wird aus der Fülle von möglichen akustischen Umgebungen die Wohn- und Büroumgebung festgelegt, da entsprechende Haushaltgeräte o. ä. in diesem Bereich platziert sind. Für die Auswahl bzw. die Entwicklung von Maßnahmen zur Robustheitssteigerung bei Hallstörungen gilt deshalb das Kriterium, dass ein praxistaugliches Verfahren bei typischen Hallbedingungen aus Wohn- und Büroumgebungen funktionieren und die-

se vollständig abdecken muss. In anderen Umgebungen wird eine Funktionsfähigkeit nicht verlangt. In welchen Grenzen sich die Hallbedingungen in diesen Umgebungen bewegen, beantwortet eine Untersuchung in Abschnitt 3.4.2. Die angesprochene Vernachlässigung des Rauschens $n(t) = 0$ für diese Arbeit ist zunächst eine Annahme, die für die betreffenden Umgebungen in weiten Grenzen gültig sein kann. Oft findet man in Büros bzw. Wohnräumen Stille vor. Auftretende stationäre Geräusche haben meist geringe Lautstärken, z. B. Geschirrspüler, Lüfter von Klimaanlagen oder Computern etc. Es gilt dennoch, dass für die Auswahl oder Entwicklung robustheitsteigernder Maßnahmen gegen Hall ein mögliches Geräusch in die Überlegungen mit einbezogen wird. So ist z. B. der in Kapitel 6 vorgestellte Ansatz prinzipiell auch unter Geräuschbedingungen einsetzbar, obwohl der praktische Nachweis dieser Behauptung den Rahmen der Arbeit übersteigt. Neben den stationären Geräuschen sind in der Realität auch instationäre Geräusche, wie Sprache konkurrierender Sprecher, Fernseh- oder Radioton, Türen- oder Fensterklappen, typische Küchengeräusche beim Kochen, Abwaschen, Essen etc., vorhanden. Die Behandlung von stationären und instationären Geräuschen wird in dieser Arbeit nicht weiter untersucht.

1.3 Einordnung der Arbeit

Versuche, die Sprachverarbeitung technisch nachzubilden, sind sehr alt. Als ein erster entscheidender Meilenstein gilt die Sprechmaschine von v. Kempelen, die 1791 publiziert wurde [BSH08b]. Unabhängig von der Sprachverarbeitung entwickeln sich ebenfalls seit dem 18. Jahrhundert die Theorien der Signale und der Systeme als eigene Wissenschaften. Dabei entsteht auch das Teilgebiet der zeitdiskreten Signale und Systeme. Mit der Erfindung des Computers[13] ist es später möglich, zeitdiskrete Signale als Zahlenkolonnen einzulesen. Die zuvor benötigte Quantisierung wandelt die analogen Werte in eine wertediskrete, vom Computer speicherbare Form um, die digitalen Signale. Die Verarbeitung digitaler Signale stellt seitdem eine eigene Wissenschaft dar, die stets mit der Signalverarbeitung kontinuierlicher und diskreter Signale in Beziehung steht. Die digitale Sprachsignalverarbeitung stellt ein besonderes Teilgebiet der digitalen Signalverarbeitung dar, das sich speziell mit Problemstellungen bei Sprachsignalen befasst. Ihre Geschichte beginnt erst mit der verbreiteten Verfügbarkeit von Computern zu Beginn der 1970er Jahre. Seitdem ist die Fülle der betreffenden Problemstellungen stetig gestiegen, sodass eine Einteilung in Teilbereiche mittlerweile sehr komplex ist. Grob lässt sich eine Einteilung wie folgt angeben:

[13] In der Geschichte von Rechenapparaturen bis hin zum Computer gibt es verschiedene Meilensteine, von denen einer der 1941 von Zuse in Betrieb genommene Z3 ist, der allgemein als erster funktionsfähiger Computer der Geschichte angesehen wird. Für die digitale Signalverarbeitung waren jedoch erst spätere Generationen von Computern von Bedeutung, da deren Algorithmen bekanntermaßen rechenintensiv sind.

- Modellierung der Prozesse der menschlichen Spracherzeugung sowie -perzeption,
- spezielle Signalverarbeitungsverfahren für Sprachsignale (z. B. Grundfrequenzdetektion, Signalanalyse, etc.),
- Sprachkodierung und -kompression,
- Sprachsynthese,
- Spracherkennung[14],
- Sprecher- und Sprachenerkennung,
- Speech Enhancement (z. B. Geräuschunterdrückung[15], akustische Echokompensation[16], Enthallung[17]) und
- räumliche Problemstellungen (z. B. blinde Quellenortung[18] und -trennung[19], Beamforming u. a.).

Diese Einteilung findet man ähnlich in [BSH08a], wo ein umfassender und gut strukturierter Überblick auf hohem Niveau zum aktuellen Stand der Wissenschaft der digitalen Sprachsignalverarbeitung gegeben wird.

Arbeiten unter Mitwirkung des Autors bewegen sich bislang auf den Gebieten der robusten Spracherkennung[20], der Geräuschunterdrückung[21], der Enthallung[22], der akustischen Echokompensation[23], Beamforming[24], Grundfrequenzdetektion[25], Signalanalyse[26] und Voice Activity Detection[27].

Die vorliegende Arbeit setzt genannte traditionelle Themen der Sprachsignalverarbeitung mit dem Gebiet der Raumakustik in Beziehung. Die Untersuchung des Verhaltens von Spracherkennung unter raumakustischen Umgebungsbedingungen sowie das Entwickeln von entsprechenden robustheitssteigernden Maßnahmen stellt ein neues Forschungsgebiet dar, das gerade einige wenige Ansätze hervorgebracht hat (vgl. Kapitel 5). Das erkennt man bereits daran, dass dieses Thema in [BSH08a] gar nicht beleuchtet wird, wohingegen Robustheit gegen Umgebungsgeräusche ein eigenes Kapitel erhält [Dro08] (ein frühes Standardwerk ist [JH95]). Erweiternd darf gesagt werden, dass der Einfluss von raumakustischen Umgebungsbedingungen auf Methoden und Verfahren aus anderen Gebieten der Sprachsignalverarbeitung ebenfalls kaum untersucht ist.

[14] Engl. Fachbegriff: Automatic Speech Recognition (ASR).
[15] Engl. Fachbegriff: Noise Reduction (NR).
[16] Engl. Fachbegriff: Acoustic Echo Cancelation (AEC).
[17] Engl. Fachbegriff: Dereverberation, Blind Dereverberation.
[18] Engl. Fachbegriff: Blind Source Localization (BSL).
[19] Engl. Fachbegriff: Blind Source Separation (BSS).
[20] Zugehörige Arbeiten sind [HBK+02, PHHJ03, PHHJ04, PHGK05, PKH05, PLWH07, WPWH07, PJH07, HAA+08, Lor08, PLU+08b, PLU+08a, PLLH08].
[21] Zugehörige Arbeiten sind [PHHJ03, PHHJ04, PGF04, PHGK05, PKH05, WPWH07].
[22] Zugehörige Arbeiten sind [Loh07, PLLH08, PLU+08a, PLU+08b].
[23] Zugehörige Arbeiten sind [Pet01, PGF04, Gru04].
[24] Zugehörige Arbeiten sind [PHGK05, PKH05].
[25] Zugehörige Arbeiten sind [PUM+08, UPMH08].
[26] Zugehörige Arbeiten sind [HBK+02, PLLH08].
[27] Zugehörige Arbeiten sind [HBK+02, RPH03, EKPH06].

Ein Beispiel dafür ist die Grundfrequenzdetektion (F_0-Detektion) und die Stimmhaft-Stimmlos-Unterscheidung (engl.: VUD – Voiced Unvoiced Decision), weshalb hier ein eigenes Kapitel (Kapitel 7) dazu vorgesehen ist. Eine Ausnahme bildet das Gebiet der blinden Enthallung, das sich seit etwa zehn Jahren speziell diesem Thema widmet. Auch räumliche Aufgabenstellungen der Sprachsignalverarbeitung, wie blinde Quellenortung und -trennung, haben das Problem des Raumhalls und werden ebenfalls seit ca. zehn Jahren dahingehend untersucht.

Aus den zahlreichen Leistungsklassen der Spracherkennung, wie z. B. der sprecherabhängigen oder -unabhängigen Erkennung fließender oder spontaner Sprache, mit großen oder kleinen Vokabularien usw., beschränkt sich diese Arbeit auf die Erkennung von Kommandowörtern für Gerätesteuerungen mit kleinen Vokabularien, woraus sich die Randbedingungen aus dem vorigen Abschnitt ergeben. Es wird versucht, die Spracherkennung auf Signal- bzw. Merkmalebene mit Lösungsansätzen und Algorithmen der Signal- und Merkmalanalyse zu verbessern. Gelegentlich spricht man dabei auch von Methoden am akustischen Front-End. Auf die eigentliche Klassifikation mit HMMs wird nicht bzw. nur soweit wie nötig eingegangen.

1.4 Überblick und wissenschaftliche Beiträge der Arbeit

Kapitel 2 – Experimentierumgebung – Überblick und Definitionen: Bevor sich die Arbeit mit der eigentlichen Thematik der robusten Spracherkennung unter raumakustischen Umgebungsbedingungen befasst, wird kurz das Experimentiersystem vorgestellt. Gleichzeitig werden anhand dessen wichtige, später benötigte Definitionen und Größen der Signalverarbeitung sowie einige für die Arbeit konstant festgelegte Parametereinstellungen eingeführt. Die Besonderheiten des im Experimentiersystem vorhandenen State-of-the-Art-Erkenners werden erwähnt. Er bildet den Ausgangspunkt, an dem sich die Verbesserungen messen, und wird deshalb zu Beginn überblicksartig eingeführt.

Kapitel 3 – Raumakustische Umgebungsbedingungen: Mit der Untersuchung raumakustischer Phänomene wird in diesem Kapitel der Grundstein für die in dieser Arbeit behandelte Thematik gelegt. Es werden für die Raumakustik relevante akustische Vorgänge beleuchtet, die mit der Wiederholung akustischer Grundlagen abgeleitet werden. Dabei spielt die Strecke Sprecher-Raum-Mikrofon (SRM) eine besondere Rolle, auf ihrem Weg wird das Sprachsignal durch den Raum (Raumhall) störend beeinflusst. Aus den akustischen Phänomenen wird eine systemtheoretische Betrachtung des Raumes abgeleitet, es entsteht das SRM-System. In einer statistischen Studie wird das raumakustische Verhalten von Wohn- und Büroräumen untersucht. Eine ähnliche statistische Untersuchung ist dem Autor nicht bekannt. Es können generelle Eigenschaften der Raumklassen im Wohn- und Büroumfeld herausgefunden werden, die den

in Kapitel 6 vorgestellten neuen Ansätzen zu Gute kommen. Das Kapitel beschreibt an sich keine wissenschaftlichen Neuheiten, allerdings ist diese detaillierte Betrachtungsweise neu im Zusammemhang mit der Spracherkennung sowie Enthallungsverfahren.

Kapitel 4 – Die Wirkung des Raumes auf Sprache und Spracherkennung: Nach der Beschreibung des Raumes wird untersucht, in welcher Weise der Raum das Sprachsignal beeinflusst. Dabei werden zunächst Eigenschaften ungestörter Sprachsignale vorgestellt. Im Anschluss wird signaltheoretisch beschrieben, wie das SRM-System auf das saubere Sprachsignal einwirkt. Bereits bestehende Maße, die die Störung durch den Raum beschreiben, werden diskutiert und kurz beurteilt. Es wird festgestellt, dass aktuelle Maße die Störung für die Spracherkennung nicht adäquat beschreiben, worauf in dieser Arbeit ein neues Maß vorgeschlagen wird. Im Anschluss wird untersucht, wie der Raum auf die Spracherkennung einwirkt. Dabei werden Abhängigkeiten von zwei wesentlichen raumakustischen Parametern nachgewiesen, die Abhängigkeit von der Nachhallzeit T_{60} sowie vom SMD. Vereinzelt wurde der SMD bereits in Experimenten anderer Wissenschaftler berücksichtigt, allerdings stellt die in dieser Arbeit eingeführte Systematik in den Untersuchungen eine Neuerung dar. Im Anschluss werden Experimente durchgeführt, bei denen nur bestimmte Teile oder Phasen des Halls im Sprachsignal belassen werden. Durch eine systematische Variation dieser Phasen liefern die Experimente neue Erkenntnisse über besonders störende und weniger störende Bereiche des Halls für die Spracherkennung. Durch diese Erkenntnisse werden unter anderem die neuen Ansätze in Kapitel 6 motiviert.

Kapitel 5 – Existierende Lösungsansätze: Nachdem in den vorhergehenden Kapiteln die Problemstellung des Raumhalls für die Spracherkennung ausgiebig beschrieben wurde, stellt dieses Kapitel bereits bestehende Lösungsansätze vor. Es beschreibt den Stand der Technik. Es wird schnell deutlich, dass es sich hierbei um ein noch junges Forschungsthema handelt. Die Kategorisierung bestehender Ansätze wird an die Einteilung für robustheitssteigernde Maßnahmen gegen Störgeräusche angelehnt. Speziell auf Signalebene besteht das Forschungsgebiet der blinden Enthallung, hier existieren die meisten Ansätze. Auf anderen Ebenen ist die Anzahl der Ansätze weitaus geringer. Eine befriedigende Lösung der Problematik unter Gesichtspunkten des praktischen Einsatzes, wie sie in Abschnitt 1.2 vorgestellt werden, existiert nicht. Die praxisrelevante Bewertung kann ebenfalls als neu angesehen werden.

Kapitel 6 – Harmonicity based Feature Analysis: Dieses Kapitel beschreibt einen im Rahmen der Arbeit neu entwickelten Ansatz der Merkmalanalyse zur Steigerung der Robustheit gegen Raumhall. Er basiert auf drei Ideen, die teilweise aus anderen Ansätzen stammen bzw. in den vorhergehenden Kapiteln entwickelt worden sind. Es wird experimentell gezeigt, dass die Spracherkennung mit diesem Ansatz, im Gegensatz zur herkömmlichen Analysemethode, bei einem Großteil der relevanten raumakustischen Störungen (Wohn- und Büroumfeld) zufriedenstellend arbeitet.

Die Experimente unterteilen sich wieder in die in Kapitel 4 herausgefundenen beiden Abhängigkeiten von T_{60} und vom SMD. HFA wird zusätzlich mit bekannten Ansätzen aus der Literatur verglichen. Dazu werden von den in Kapitel 5 vorgestellten Ansätzen, von denen in Hinblick der praktischen Einsetzbarkeit nur wenige in Frage kommen, drei Methoden ausgewählt. Im Vergleich zeigt sich, dass HFA gleich gut und teilweise besser arbeitet. Besondere Vorteile von HFA ergeben sich in Hinblick der Implementierung unter Ressourcenknappheit.

Kapitel 7 – F_0-Detektion und VUD unter verhallten Umgebungsbedingungen: Die in Kapitel 6 vorgeschlagene Methode benutzt u. a. zwei Basistechnologien in der Sprachsignalverarbeitung, die F_0-Detektion und die VUD. Deshalb wird in diesem Kapitel über eine detaillierte Untersuchung des Verhaltens beider Technologien unter raumakustischen Störbedingungen berichtet. Nach der Vorstellung einer Evaluationsumgebung sowie einem Überblick über existierende Verfahren für die beiden Technologien zeigen Experimente jeweils erwartungsgemäß, dass sie wesentlich schlechter als im ungestörten Fall funktionieren. Die Experimente unterteilen sich wie zuvor in die Abhängigkeiten von T_{60} und vom SMD. Die Untersuchung beider Technologien unter raumakustischen Umgebungsbedingungen stellt eine Neuerung dar. Im Ergebnis werden Schlussfolgerungen gezogen, auf deren Grundlage je ein günstig arbeitendes Verfahren ausgewählt wird. Der VUD-Ansatz wurde dabei in dieser Arbeit neu vorgestellt. Sowohl bei der Auswahl als auch bei der Entwicklung wurde besonders auf den Ressourcenverbrauch Wert gelegt.

Kapitel 8 – Zusammenfassung und Ausblick: In diesem Kapitel wird ein Überblick über die Arbeit gegeben und deren wissenschaftliche Beiträge zusammengefasst. Es wird festgestellt, dass das Thema noch viel Potential für zukünftige Forschung bietet, da die Untersuchungen erst am Anfang stehen.

2 Experimentierumgebung – Überblick und Definitionen

2.1 Überblick

Um die Auswirkungen von realen akustischen Umgebungsbedingungen auf Spracherkennungssysteme zu untersuchen, ist es nötig, eine Experimentierumgebung zu schaffen. Sie besteht aus einem Spracherkenner, Datenbasen, der Simulation verschiedener Umgebungsbedingungen sowie Methoden der Evaluation der Erkennungsleistung. Dieser Abschnitt beschreibt die verwendete Experimentierumgebung. Dabei werden zunächst der verwendete Spracherkenner überblicksartig vorgestellt und wichtige, damit zusammenhängende Größen definiert. Im Anschluss werden die Evaluation von Spracherkennern beschrieben und ein Testset für die Experimente dieser Arbeit festgelegt.

2.2 Spracherkennungssystem

Für Spracherkennungsexperimente steht ein bereits existierendes Laborsystem zur Verfügung. Es verfolgt den Ansatz der Spracherkennung und -synthese mit gemeinsamen Datenbasen und wird deshalb mit UASR (engl.: Unified Approach for Speech Synthesis and Recognition) bezeichnet. Das UASR-System wird in [EWH00, HEW07] vorgestellt. Detaillierte Informationen können den Dissertationen von Eichner [Eic07] bzw. Werner [Wer08] entnommen werden. Für die Spracherkennungsexperimente wird der Erkennerzweig des UASR-Systems in seiner Anwendung als Kommandoworterkenner benutzt. Er bildet den Ausgangspunkt der experimentellen Forschungsaktivitäten dieser Arbeit; bisher sind keine Maßnahmen zur Robustheitssteigerung unter realen akustischen Umgebungsbedingungen implementiert.

Der schematische Aufbau des Kommandoworterkenners ist in Abbildung 2.1 dargestellt. Er unterteilt sich in die Blöcke

- primäre Merkmalanalyse,
- sekundäre Merkmalanalyse und
- Klassifikator.

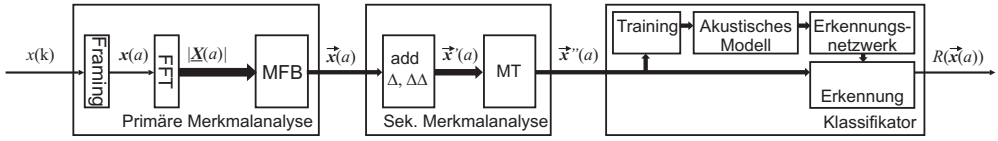

Abbildung 2.1 – Allgemeines Blockschaltbild eines Kommandoworterkenners.

Die primäre Merkmalanalyse wird im Folgenden genauer beschrieben, da hier die Front-End-Verarbeitung stattfindet und die in dieser Arbeit untersuchten Methoden integriert werden. Die sekundäre Merkmalanalyse sowie der Klassifikator werden in dieser Arbeit nicht untersucht. Beide werden deshalb nur grob beschrieben und gelten als konstant; ihre Algorithmen und ihre Parametrierung bleiben in den Experimenten unverändert.

2.2.1 Primäre Merkmalanalyse

Der Klassifikator eines Spracherkenners arbeitet nicht mit dem Zeitsignal, sondern mit Merkmalvektorfolgen, die aus dem Zeitsignal $x(t)$ extrahiert werden. Das Zeitsignal liegt dabei als kontinuierliche Folge von Abtastwerten $x(k)$ (engl.: Samples) vor. Die Abtastrate beträgt für alle in dieser Arbeit behandelten Zeitsignale

$$f_s = 16000 \text{ Hz}. \tag{2.1}$$

Aus $x(k)$ werden Kurzzeit-Signalabschnitte (engl.: Frame) ausgeschnitten und auf bestimmte Merkmale hin analysiert. Die Kurzzeit-Signalanalyse entspricht einer Datentransformation, die die Datenmenge reduziert, Spezifika der Sprecher eliminiert und sprecherunspezifische, aber lautspezifische Merkmale extrahiert. Aus jedem Frame a wird ein Merkmalvektor $\vec{\boldsymbol{x}}(a)$ generiert, es entsteht die oben genannte Merkmalvektorfolge.

Framing Das Framing stellt aus dem Eingangssignal $x(k)$ der Länge K die gerade erwähnte Folge von Kurzzeit-Signalabschnitten $\boldsymbol{x}(a)$ der Länge K' zur Verfügung. Für die Sprachsignalanalyse überlappen sich die Frames, wobei die Bandbreite der benutzten, nachfolgend beschriebenen Fensterfunktion als Ausgangsgröße zur Bestimmung des Fortsatzintervalls FI dient (Grundlage ist das Abtasttheorem für die Kurzzeit-Signalanalyse.). D. h., mit dem Fortsatzintervall $FI < K'$ wird durch Hochzählen des Frameindex a das Analysefenster vorwärts bewegt. Die Analyseframes können auch als Matrix $x'(a, k')$ geschrieben werden

$$\boldsymbol{x}(a) = x'(a, k') = x(a \cdot FI + k') \; ; \quad 0 \le k' \le K' - 1 \; ; \quad 0 \le a \le A - 1. \tag{2.2}$$

k' ist der frameinterne Abtastindex. Für dateibasierte Verarbeitung existiert für a die Obergrenze A, die sich aus der Signallänge K sowie FI berechnet[1]. Da diese Arbeit im Folgenden keine Berührung mehr mit einem kontinuierlichen Zeitsignal in Zusammenhang mit einer Signalanalyse hat, wird an dieser Stelle zur Vereinfachung die Notation der Matrix in $x(a,k)$ geändert

$$\boldsymbol{x}(a) = x'(a,k') = x(a,k)\,; \quad 0 \leq k \leq K-1\,; \quad 0 \leq k' \leq K'-1\,; \quad 0 \leq a \leq A-1, \tag{2.3}$$

sodass aus k' der frameinterne Index k und aus K' die Framelänge K wird. In der benutzten Experimentierumgebung gelten

$$\begin{array}{rclcrcl} K & = & 400 & \rightarrow & T_K & = & 25 \text{ ms} \\ FI & = & 160 & \rightarrow & T_{FI} & = & 10 \text{ ms}. \end{array} \tag{2.4}$$

Für die framebasierten Algorithmen zur F_0-Detektion in Kapitel 7 gelten die gleichen Definitionen für das Framing, die Werte für FI und K können jedoch unterschiedlich besetzt sein.

Spektralanalyse Für jeden Frame $\boldsymbol{x}(a)$ wird eine spektrale Analyse durchgeführt. Für den Rest dieses Abschnitts soll deshalb nur der aktuelle Frame beleuchtet werden; aus Vereinfachungsgründen entfällt daher hier der Frameindex a. Die Spektralanalyse besteht aus Preemphase (die im Rahmen dieser Arbeit jedoch ausgeschaltet ist), Fensterung (es sind verschiedene Fensterfunktionen implementiert, benutzt wird ein Hammingfenster [OS04]), Zeropadding mit $N - K = 112$ Nullen und FFT[2] ($N = 512 \rightarrow T_N = 32$ ms). Am Ausgang der FFT, der schnell arbeitenden Implementierung der DFT[3]

$$\underline{X}(n) = \frac{1}{N} \sum_{k=0}^{N} x(k) e^{-j2\pi \frac{nk}{N}}, \tag{2.5}$$

entsteht das komplexe Spektrum $\underline{X}(n)$ mit dem Frequenzindex n. Die Phase spielt in der weiteren Verarbeitung keine Rolle, sodass entweder mit dem Amplitudenspektrum

$$A(n) = |\underline{X}(n)| \tag{2.6}$$

[1] Man kann sich auch ein Onlinesystem vorstellen, das aufgrund des einlaufenden Datenstroms keinen Abschluss besitzt. Als Onlinesystem wird hier ein System verstanden, das einen ständig eingeschalteten Signaleingang hat, etwa bei einem dauernd eingeschalteten Mikrofon. In der Realität gibt es allerdings keine unendliche Einschaltzeit. Es existiert also auch hier ein A bzw. K. Da es allerdings technisch keinen Sinn hat, unendlich (bzw. derart) große Matrizen $x(a,k)$ zwischenzuspeichern, muss eine Automatik implementiert werden, die das Signal sinnvoll in Blöcke zerlegt. Ein Beispiel dafür ist ein VAD, der genau ein Wort ausschneidet, für das dann a, A, k und K gilt.

[2] FFT – Fast Fourier Transform.

[3] DFT – Diskrete Fourier Transformation.

(in dieser Arbeit mit $|\underline{X}(n)|$ bezeichnet) oder mit dem Leistungsspektrum

$$S(n) = |\underline{X}(n)|^2 = \underline{X}(n) \cdot \underline{X}^*(n) \tag{2.7}$$

gerechnet wird. Der Frequenzschritt zwischen zwei Frequenzindizes n und $n+1$ der hier benutzten FFT beträgt

$$\Delta f = \frac{f_s}{N} = 31,25 \text{ Hz}. \tag{2.8}$$

Für die Spracherkennung wird in dieser Arbeit das Amplitudenspektrum verwendet. Genauere Informationen zur Signalanalyse finden sich z. B. in [Hof98].

Merkmalextraktion Die Merkmalextraktion extrahiert aus der noch relativ großen Datenmenge des positiven Frequenzbereichs von $|\underline{X}(n)|$ (hier 256 Werte) einen geringer dimensionalen Merkmalvektor $\vec{\boldsymbol{x}}(a)$. Die Merkmale sollen dabei möglichst nicht sprecherspezifisch, dafür aber lautspezifisch sein. Im Wesentlichen bedeutet dies, dass in stimmhaften Abschnitten die variable Grundfrequenz eliminiert wird und die lautspezifischen Formantinformationen erhalten bleiben (vgl. Abschnitt 4.2.1). Um dies zu erreichen, existieren in der Sprachverarbeitung typische Merkmalextraktionsverfahren [RJ93]. Die zwei wichtigsten Verfahren sind:

- **MFB – Mel-Filterbank:** Die Mel-Filterbank besteht aus einer Gruppe von C Bandpassfiltern, deren Mittenfrequenzen auf einer Mel-Skala

  $$Z = 2595 \cdot \lg\left(1 + \frac{f}{700 \text{ Hz}}\right) \tag{2.9}$$

  (Gleichung z. B. in [Rud99]) äquidistant verteilt und deren Bandbreiten im Mel-Bereich gleich groß sind [Hof98, VHH98]. Im Frequenzbereich bedeutet dies, dass sowohl die Abstände ihrer Mittenfrequenzen als auch ihre Bandbreiten mit steigender Frequenz größer werden. Am Ausgang des Filterkanals c wird die Energie $e_{c,\text{MFB}}$ als Element eines Energievektors $\vec{\boldsymbol{e}}_{\text{MFB}}$ gebildet. Es entsteht ein Vektor, der einem stark unterabgetasteten, geglätteten Leistungsspektrum mit Mel-verzerrter Frequenzachse entspricht. Um die große Dynamik von Sprachsignalen zu berücksichtigen, wird die Energie logarithmiert und in dB angegeben

  $$\vec{\boldsymbol{x}}_{\text{MFB}} = 10 \lg \vec{\boldsymbol{e}}_{\text{MFB}} \text{ dB}, \tag{2.10}$$

  was letztlich die typische Definition eines MFB-Merkmalvektors darstellt.

- **MFCC – Mel-Frequenz-Cepstral-Koeffizienten:** MFCCs bilden das Cepstrum aus einem Spektrum, dessen Frequenzen auf eine Mel-Skala abgebildet werden (Mel-Spektrum) und das dadurch entsprechend verzerrt erscheint. Zur Bildung des Cepstrums wird das Mel-Spektrum logarithmiert und im Anschluss

mit einer diskreten Kosinustransformation (DCT) [ANR74] cepstral analysiert. Die für den Laut charakteristische Grundform ist niederquefrent, sodass die unteren Koeffizienten zur Beschreibung der Quefrenzstruktur ausreichend sind. Höhere Quefrenzen sind nicht relevant; hier wird bspw. die sprecherspezifische Grundfrequenz abgebildet. Eine übliche Methode, MFCCs zu erzeugen, besteht in der Benutzung eines MFB-Merkmalvektors anstelle des Mel-Spektrums, auf den die DCT

$$\vec{\boldsymbol{x}}_{\mathrm{MFCC}} = \mathrm{DCT}(\vec{\boldsymbol{x}}_{\mathrm{MFB}}) \tag{2.11}$$

$$x_{q,\mathrm{MFCC}} = \sum_{c=0}^{C-1} x_{c,\mathrm{MFB}} \cdot \cos\left(\frac{\pi}{C}\left(c+\frac{1}{2}\right)q\right) \;;\; q = 0,\ \ldots,\ C-1 \tag{2.12}$$

angewendet wird. Es wird noch einmal betont, dass die MFB-Vektoren logarithmiert vorliegen müssen (Gleichung (2.10)), da man sonst nicht in den Cepstralbereich gelangt. Der MFB-Vektor, der im Vergleich zum Mel-Spektrum eine wesentlich geringere Dimension besitzt, bildet die niederquefrente Grundform des Mel-Spektrums bereits ab; er stellt eine Unterabtastung des geglätteten Mel-Spektrums dar. Eine Höherabtastung ist nicht nötig, da sich dadurch nur die nicht benötigten hochquefrenten Anteile herausbilden, die im Ergebnis der DCT wieder eliminiert werden würden.

Spracherkennungssysteme sind individuelle Systeme, abgesehen von Grundprinzipien weichen Einzelheiten in den Algorithmen sowie Parameter in den unterschiedlichen Implementierungen oft voneinander ab. Für verteilte Systeme, wo Endgeräte verschiedener Nutzer Merkmalvektoren über ein Netz senden und diese von einem Spracherkennungsserver verarbeitet werden, existiert eine notwendige, allgemein anerkannte Norm für die primäre Merkmalextraktion [ETS03a, ETS03b, ETS05a, ETS05b, MMN+02]. Darin werden 12-dimensionale MFCC-Vektoren als Standard definiert. Das hier benutzte Experimentiersystem weicht davon ab. Es arbeitet standardmäßig mit MFB-Vektoren, die einer 30-dimensionalen Mel-ähnlichen Skala folgen [Eic07]. Im Unterschied zur oben beschriebenen MFB werden im Vorfeld die Spektren cepstral geglättet. Die einzelnen Vektoren werden auf die Energie des aktuellen Frames normiert (Subtraktion im logarithmischen Bereich), die zusätzlich das 31. Merkmal im Merkmalvektor bildet. Die genaue Implementierung ist in [Wes96] nachzulesen.

2.2.2 Sekundäre Merkmalanalyse

Die sekundäre Merkmalanalyse besteht aus zwei Schritten, der Bildung kontextabhängiger Merkmale und einer Merkmaltransformation.

Kontextabhängige Merkmale Kontextabhängige Merkmale bilden aus der Differenz zu Vorgängern und Nachfolgern des aktuellen Merkmalvektors $\vec{\boldsymbol{x}}(a)$ sogenannte dy-

namische Merkmale, die erstmals von Furui [Fur86] untersucht wurden. Dabei gibt es verschiedene Ansätze [ST99], von denen hier die Differenzen erster und zweiter Ordnung benutzt werden [Wes96, Eic07]. Sie werden auch als Δ- und ΔΔ-Merkmale bezeichnet. Der MFB-Vektor wird um die Δ- und ΔΔ-Merkmale erweitert, sodass ein 93-dimensionaler Supervektor $\vec{x}'(a)$ entsteht.

Merkmaltransformation Die Merkmaltransformation (MT) hat die Intention, Redundanzen, Korrelationen und irrelevante Informationen aus den Merkmalvektoren zu eliminieren. Es existieren verschiedene Verfahren der Merkmaltransformation, von denen hier die Hauptkomponentenanalyse (engl.: PCA – Principal Component Analysis) benutzt wird. Dabei handelt es sich um eine Multiplikation mit einer Transformationsmatrix $\mathbf{W}$

$$\vec{x}''(a) = \mathbf{W} \cdot (\vec{x}'(a) - \overline{x}) \tag{2.13}$$

mit $\vec{x}'(a)$, der zuvor noch von den statistischen Mittelwerten $\overline{x}$ befreit werden muss. $\mathbf{W}$ ist eine Matrix, bestehend aus Eigenvektoren einer Kovarianzmatrix, die aus einer signifikanten Stichprobe des Trainingsmaterials gebildet wird. $\overline{x}$ beschreibt einen Vektor aus Mittelwerten, die ebenfalls aus einer Trainingsstichprobe gebildet werden. Sowohl $\mathbf{W}$ als auch $\overline{x}$ werden nur zum Zeitpunkt des Trainings bestimmt und bleiben während der Erkennungsphase konstant. Die PCA reduziert die Vektordimension von 93 auf 24. Eine allgemeine Schreibweise der Merkmaltransformation findet man in [TW09].

2.2.3 Klassifikator

Akustisches Phonemmodell Im UASR-System ist ein Phonem-basierter HMM[4]-Klassifikator implementiert. Er besitzt ein akustisches Modell, dass aus 44 monophonen Phonem-HMMs besteht. Die einzelnen Phonem-HMMs setzen sich aus drei Zuständen zusammen (ausgenommen Anfangs- und Endzustand). Da sich die Verteilung der Realisierungen der Merkmalvektoren einzelner Zustände im Merkmalvektorraum nicht immer nach einer einzigen Gaußverteilung ausprägt, ist es üblich, sogenannte Mischverteilungen (engl.: Mixtures) zu benutzen. Darunter versteht man, dass die Realisierungen einer Trainingsstichprobe eines Zustandes in m Vektorcluster aufgeteilt werden, aus denen je eine Gaußverteilung berechnet wird. Die einzelnen m Gaußverteilungen werden gewichtet addiert. Es entsteht ein Mixturmodell (engl.: GMM – Gaussian Mixture Model). Das UASR-System weicht von diesem Standard ab. Es bildet keine GMMs mit typischer Links-Rechts-Struktur, sondern benutzt anstelle dessen m parallele Zustände mit einer einzigen Gaußverteilung, einem sogenannten stochastischen Markovgraphen

[4] HMM – Hidden Markov Modell [Rab89].

(SMG[5]). [Eic07] beschreibt, dass SMGs und GMMs ineinander umgerechnet werden können, dass jedoch SMGs gegenüber GMMs verschiedene Vorteile besitzen. Durch Versäuberungsstrategien können u. a. nicht benötigte Zustände entfernt werden, was die Komplexität des Modells vereinfacht sowie gegenüber klassischen GMMs Rechenleistung einspart. Für die Experimente in dieser Arbeit werden durchgängig SMGs mit $m = 2$ Gaußverteilungen pro Zustand verwendet[6].

Training Die Phonem-HMMs werden aus einer relevanten Menge an Sprachdaten trainiert. Als Trainingsstichprobe steht die Verbmobil-Datenbasis [Wah00] zur Verfügung. Sie besteht aus ca. 47 h Sprache. Zur Beschleunigung der Experimente wurde für diese Arbeit beim Training durchgängig mit einem Subset von ca 3,5 h Sprache gearbeitet. Das Trainingsmaterial liegt in etikettierter Form vor, d. h., die Signalabschnitte besitzen Markierungen, die das aktuelle Phonem beschreiben. Somit können die Signalabschnitte aller Phoneme einer Klasse zum Training des betreffenden Phonemmodells extrahiert werden.

Erkennung In dieser Arbeit wird der UASR-Erkenner in der Form eines Kommandoworterkenners [Eic07] benutzt. D. h., das Eingangssignal des Klassifikators ist eine abgeschlossene Merkmalmatrix eines Wortes und das Erkennungsnetzwerk besteht aus Wortmodellen[7]. Das Erkennungsnetzwerk der Wortmodelle wird aufgrund eines gegebenen Lexikons aus dem Zusammenschalten der einzelnen Phonemmodelle mit anschließender Netzwerkoptimierung und -versäuberung gebildet. Das Lexikon beschreibt die einzelnen Wortklassen des Vokabulars in einer Phonemschreibweise. In der Erkennungsphase werden die Emissionswahrscheinlichkeiten der einzelnen Zustände berechnet und eine Viterbi-Suche ermittelt den optimalen Weg durch das Erkennungsnetzwerk, woraus sich das Erkennungsergebnis $R(\vec{\boldsymbol{x}}'')$ ergibt. Eine detaillierte Beschreibung der Erkennung ist in [Eic07] nachzulesen.

2.3 Evaluation

Unter der Evaluation von Spracherkennern versteht man die Beurteilung ihrer Qualität aufgrund von Gütekriterien, u. a. Erkennungsgenauigkeit[8], benötigte Rechenleistung,

[5] Ein SMG ist ein Spezialfall eines HMMs. Die beiden Kürzel werden daher im Folgenden synonym gebraucht, wobei HMM im Allgemeinen bevorzugt wird.

[6] Genauer gesagt wurde das Modell `1_1_6` ausgewählt, das aufgrund von Anfangsexperimenten die besten Ergebnisse bei zwei Gaußverteilungen pro Zustand erzielte (zur Bedeutung des Kürzels `1_1_6` vgl. Dissertation [Eic07]).

[7] Genau genommen handelt es sich um Modelle von kurzen Kommandophrasen, vgl. Abschnitt 2.3.6.

[8] Bemerkung: Es ist nicht möglich, die Erkennungsgenauigkeit von Spracherkennern allgemein zu messen. Das liegt u. a. daran, dass die Variabilität der menschlichen Sprache sehr groß ist und

Reaktionszeit etc. Für diese Arbeit soll sich die Evaluation auf die Erkennungsgenauigkeit von Kommandoworterkennern beziehen.

2.3.1 Klassifikationsarten eines Kommandoworterkenners

Wortschatz Für eine bestimmte Erkennungsaufgabe wird ein Wortvokabular $W = (W_1;\ W_2;\ \ldots;\ W_K)$ festgelegt, das sich in K Klassen aufteilt.

Mögliche Testklassen Die möglichen (gültigen) Eingabewörter (T für Testwort) haben die Klassen $T = (T_1;\ T_2;\ \ldots;\ T_K)$, die mit W korrespondieren. Zusätzlich treten in realen Umgebungsbedingungen Wörter bzw. andere akustische Ereignisse auf, die nicht zum Vokabular gehören, sogenannte Out-of-Vocabulary-Wörter [BSH08a] (OOV-Wörter, Index O). Diese können abstrakt in einer Testklasse $T_O \notin W$ zusammengefasst werden.

Erkennungsergebnisse ohne Rückweisung Der Erkenner besitzt demnach eine Menge $R = (R_1;\ R_2;\ \ldots;\ R_K)$ von möglichen Erkennungsergebnissen (R für engl.: Recognition Result), die ebenfalls mit W korrespondieren. Beim Erkennungsvorgang wird vom Klassifikator jeder Klasse R_k ein Score zugewiesen; die Klasse mit dem Extremscore gewinnt. Die Art des Scores variiert zwischen den Erkennertypen, z. B. maximale Emissionswahrscheinlichkeit beim HMM-Erkenner oder minimale Distanz beim DTW[9]-Erkenner.

Rückweisung Unter Rückweisung (engl.: Rejection, Index Rej) versteht man, dass ein Erkenner bei einem eingegebenen Wort kein Ergebnis liefert, es zurückweist. Für das Erkennungsergebnis Rückweisung wird das Symbol R_{Rej} benutzt. Der Nutzer sieht dieses Ergebnis normalerweise nicht, das Gerät reagiert einfach nicht. Die Rückweisung unterscheidet zwei Fälle:

- $T_k \in W$: Ist der Klassifikator in der Lage, ein Maß für die Sicherheit seiner Erkennung, die Konfidenz, zu bestimmen, so soll das T_k bei Erreichen einer zu großen Unsicherheit (zu kleinen Konfidenz) zurückgewiesen werden. Die Rückweisung ist u. U. günstiger als ein unsicheres richtiges oder ein falsches Ergebnis auszugeben. Bsp. Jalousiesteuerung: der Mensch empfindet es unangenehm,

Erkenner einen bestimmten begrenzten Wortschatz besitzen, der jedoch von Erkenner zu Erkenner unterschiedlich ist/sein kann. Es ist daher nicht ohne Weiteres möglich, ein allgemeingültiges Testszenario zu erstellen. Es hat wenig Sinn, eine Erkennungsrate von 95 % anzugeben, ohne zu spezifizieren, unter welchen Bedingungen, bei welchem Testkorpus, Wortschatz oder welchen akustischen Umgebungsbedingungen die Rate gemessen worden ist. Für einen bestimmten Wortschatz und ein festgelegtes Testszenario lassen sich Erkenner untereinander vergleichen. Zusätzlich ist die Berechnungsvorschrift der Erkennungsrate nicht immer einheitlich und demnach zu spezifizieren.

[9] DTW – Dynamic Time Warping (z. B. in [Rud99]).

wenn sich die Jalousie beim Kommando *Jalousie senken* hebt, wohingegen er das Wiederholen des Kommandos in Kauf nimmt. Letzteres entspricht der normalen zwischenmenschlichen Kommunikation, bei der ein Gesprächspartner intuitiv das Gesagte wiederholt, wenn der andere es nicht verstanden hat. Typische Rückweisungsstrategien sind hier die Benutzung von Konfidenzmaßen, z. B. die Größe der Differenz zwischen den Scores von R_k und $R_{\neq k}$.

- $T_O \notin W$: OOV-Wörter müssen vom Erkenner als solche klassifiziert werden. Die Rückweisung von OOV-Wörtern ist vermutlich die schwierigste Aufgabe von Erkennungssystemen, denn im Jalousiebeispiel würde es der Anwender *nicht* akzeptieren, wenn sich die Jalousie plötzlich senkt, während der Fernseher Ton aussendet. Deshalb soll die Rückweisung möglichst fehlerfrei stattfinden, was eine anspruchsvolle Aufgabe darstellt, die derzeit kaum lösbar erscheint. Eine typische Rückweisungsstrategie für OOV-Wörter ist ebenfalls die Benutzung von Scores, d. h., wenn der Score von R_k nicht groß genug ist (es ist unwahrscheinlich, dass das Wort zum Vokabular gehört), wird zurückgewiesen. Eine sehr wirkungsvolle Strategie ist die Einführung einer sogenannten Müllklasse (oder Müllmodell, engl.: Garbage Model), die im Idealfall die komplette akustische, nicht zum Vokabular gehörende Welt abbildet. Naturgemäß ist es nicht möglich, eine solche Müllklasse vollständig zu modellieren.

Möglichkeiten der Klassifikation Aufgrund der zuvor geschilderten Konfiguration des Klassifikators ergeben sich folgende Möglichkeiten der Klassifikation:

- Eingabe $T_k \in W$
  - $T_k \rightarrow R_k$: $C_{k,k}$ korrekte Klassifikation,
  - $T_k \rightarrow R_{\neq k}$: $S_{k,i}$ falsche Klassifikation, Verwechslung (engl.: Substitution),
  - $T_k \rightarrow R_{Rej}$: $D_{k,Rej}$ Rückweisung, Auslassung (engl.: Deletion). Es kann noch genauer unterschieden werden, ob der beste Score vor dem Rückweisungsalgorithmus zu R_k (Falschrückweisung, eine korrekte Klassifikation $C_{k,k}$ wird durch Rückweisungsstrategien als Fehlklassifikation bewertet) oder zu $R_{\neq k}$ (korrekte Rückweisung, eine falsche Klassifikation $S_{k,i}$ wird durch Rückweisungsstrategien als solche bewertet) gehört. Daraus kann mit einer entsprechenden Statistik die Güte der Rückweisung ermittelt werden.
- Eingabe $T_O \notin W$
  - $T_O \rightarrow R_k$: $I_{O,k}$ Einfügung (engl.: Insertion), Falschakzeptanz, entspricht einem Fehler,
  - $T_O \rightarrow R_{Rej}$: $C_{O,Rej}$ korrekte Rückweisung von OOV-Wörtern, entspricht einem korrekten Klassifikationsergebnis.

2.3.2 Berechnung von Häufigkeiten, Verwechslungsmatrix

Für die Evaluation ist es nötig, die Erkennung über eine Teststichprobe durchzuführen. Die dabei auftretenden Einzelerkennungen werden summiert und als Häufigkeiten der Einzelerkennungen in Tabelle 2.1 aufgetragen, die die sogenannte Verwechslungsmatrix (engl.: Confusion Matrix) aufspannt. Es werden die im vorigen Abschnitt eingeführten Symbole $C_{k,k}$, $S_{k,i}$, $D_{k,Rej}$, $I_{O,k}$ in die Felder der Verwechslungsmatrix eingetragen, um die absoluten Häufigkeiten für die entsprechende Klassifikationsart auszudrücken. Maße für die Erkennungsgenauigkeit, wie Erkennungsraten oder Fehlerraten, können daraus klassenspezifisch oder global berechnet werden. Für beide Fälle sind in Tabelle 2.1 die Berechnungen für die Gesamthäufigkeiten N, D, C und S angegeben. Zur Berechnung der globalen Erkennungsrate werden die folgenden globalen Häufigkeiten benötigt:

- Anzahl korrekte Erkennungen

$$C = \sum_{k=1}^{K} C_{k,k} = N - S - D, \tag{2.14}$$

- Anzahl Verwechslungen (engl.: Substitutions)

$$S = \sum_{i=1}^{K} \sum_{k=1}^{K} S_{k,i} = N - C - D, \tag{2.15}$$

- Anzahl Auslassungen (engl.: Deletions)

$$D = \sum_{k=1}^{K} D_{k,Rej} = N - C - S, \tag{2.16}$$

- Anzahl Einfügungen (engl.: Insertions)

$$I = \sum_{k=1}^{K} I_k = N_O - C_{O,Rej}, \tag{2.17}$$

- Anzahl korrekte Rückweisungen

$$C_{O,Rej} = N_O - I. \tag{2.18}$$

Tabelle 2.1 weist darauf hin, dass sich die Evaluation des Erkenners prinzipiell in die beiden Einzelmessungen Erkennungstest und Insertion-Test unterteilt. Beide stellen unterschiedliche Aufgaben für den Erkenner dar, es werden unterschiedliche Testkorpora benötigt.

Tabelle 2.1 – Vollständige Darstellung der Möglichkeiten eines Erkennungslaufes. Die unterste Zeile zeigt die Berechnung der globalen Werte für einen Erkennungstest (Insertion-Test ist ausgeschlossen.).

N_k		R_1	R_2	$\cdots$	R_K	R_{Rej}	C_k	S_k	RR_k
N_1	T_1	$C_{1,1}$	$S_{1,2}$	$\cdots$	$S_{1,K}$	$D_{1,Rej}$	$C_{1,1}$	$\sum_{k=1}^{K} S_{1,k}$	$RR_1 = \frac{C_{1,1}}{N_1}$
N_2	T_2	$S_{2,1}$	$C_{2,2}$	$\cdots$	$S_{2,K}$	$D_{2,Rej}$	$C_{2,2}$	$\sum_{k=1}^{K} S_{2,k}$	$RR_2 = \frac{C_{2,2}}{N_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N_K	T_K	$S_{K,1}$	$S_{K,2}$	$\cdots$	$C_{K,K}$	$D_{K,Rej}$	$C_{K,K}$	$\sum_{k=1}^{K} S_{K,k}$	$RR_K = \frac{C_{K,K}}{N_K}$
N_O	T_O	$I_{O,1}$	$I_{O,2}$	$\cdots$	$I_{O,K}$	$C_{O,Rej}$			$RR_O = \frac{C_{O,Rej}}{N_O}$
ges. $N = \sum_{k=1}^{K} N_k$						$D = \sum_{k=1}^{K} D_{k,Rej}$	$C = \sum_{k=1}^{K} C_{k,k}$	$S = \sum_{i=1}^{K}\sum_{k=1}^{K} S_{i,k}$	$RR = \frac{C}{N}$

2.3.3 Erkennungstest

Um die Güte des Erkenners zu evaluieren, wird eine Teststichprobe mit $N = (N_1;\ N_2;\ \ldots;\ N_K)$ Realisierungen festgelegt, deren Klassenzugehörigkeiten $T = (T_1;\ T_2;\ \ldots;\ T_K) \in W$ bekannt sind. Die Erkennungsrate RR (engl.: Recognition Rate) des Erkenners wird also zunächst mit

$$\mathrm{RR} = \frac{C}{N} = \frac{N - S - D}{N} \tag{2.19}$$

und die Fehlerrate ER (engl.: Error Rate) mit

$$\mathrm{ER} = \frac{S}{N} = \frac{N - C - D}{N} \tag{2.20}$$

gemessen[10]. Man erkennt, wie die Rückweisung die Fehlerrate mindert. Eine Rückweisungsrate RejR (engl.: Rejection Rate)

$$\mathrm{RejR} = \frac{D}{N} \tag{2.21}$$

kann ebenfalls gemessen werden. RejR ist als solche kein Qualitätsparameter eines Erkenners, allerdings lässt sich darüber die Rückweisungsschwelle optimieren, wenn RejR im Zusammenhang mit ER und RR betrachtet wird.

[10] Es ist eine Definitionsfrage, ob die Deletions D zur Fehlerrate hinzugezogen werden. In Arbeiten des Autors werden sie es nicht.

2.3.4 Insertion-Test

Hierbei wird die Rückweisung von OOV-Wörtern getestet. Dazu ist ein Testkorpus nötig, das eine Anzahl N_O von OOV-Wörtern enthält (Klasse T_O). Der Erkenner soll die Eingabeereignisse möglichst vollständig auf R_{Rej} klassifizieren. Es entsteht die OOV-Rückweisungsrate

$$\mathrm{RR}_O = \frac{C_{O,Rej}}{N_O}, \tag{2.22}$$

die möglichst nahe an 100 % liegen sollte. Im Umkehrschluss stellt die Insertion Rate

$$\mathrm{IR} = \frac{I}{N_O} \tag{2.23}$$

eine Fehlerrate dar, die möglichst gegen 0 streben sollte. Es ist auch möglich, eine IR bezogen auf die Zeit zu messen

$$\mathrm{IR}_t = \frac{I}{t}. \tag{2.24}$$

Dies geschieht vor dem Hintergrund, dass bei einem Erkenner mit einem ständig mithörenden Mikrofon möglichst wenige Insertions pro Tag auftreten sollen[11]. Die Messung von IR_t ist allerdings nur sinnvoll, wenn eine definierte OOV-Testreihe läuft, wie es bspw. beim Abspielen einer identischen Radioaufzeichnung erfolgt. Ein solcher Test wurde in der praktisch orientierten APOLLO-Evaluation [MHK+03] durchgeführt. Diese Verfahrensweise hat den Vorteil, dass die OOV-Ereignisse zufällig generiert werden, da die vollständige akustische OOV-Welt nicht nachgebildet werden kann. Allerdings hat sie den Nachteil, dass IR_t einen Wert ergibt, der nur in einer ständig beschallten Umgebung relevant ist (Extrembedingung); Stille in der Nacht etc. wird dabei nicht berücksichtigt. Dadurch erscheint der IR_t-Wert höher als er in der Praxis tatsächlich entsteht. Ein Verfahren nach Gleichung (2.23) mit einem definierten OOV-Testkorpus würde der Autor vorziehen.

2.3.5 Kombinierte Messungen

In der Literatur wird meist die sogenannte Wortfehlerrate WER (engl.: Word Error Rate)

$$\mathrm{WER} = \frac{S + D + I}{N} \tag{2.25}$$

[11] Hier erkennt man bereits die oben angesprochene Schwierigkeit der Rückweisung, da ein Nutzer (wieder Bsp. Jalousiesteuerung) ein spontanes, nicht initiiertes Heben oder Senken der Jalousie nicht tolerieren würde. Angenommen er tolerierte es einmal pro Woche ($\mathrm{IR}_t = \frac{1}{168\ \mathrm{h}}$), dann bedeutet das bei einer durchschnittlichen Wortdauer von 1 s sowie einem ständig laufenden Radiogeräusch, dass $\mathrm{IR} = \frac{1\ \mathrm{s}}{7 \cdot 24 \cdot 3600\ \mathrm{s}} 100\ \% = 0,00016\ \%$ bzw. eine gedachte $\mathrm{RR}_O = 99,99983\ \%$ beträgt. Diese extrem hohen Anforderungen sind derzeit nicht ohne Weiteres realisierbar.

bzw. die Worterkennungsrate WRR (engl.: Word Recognition Rate)

$$\mathrm{WRR} = 1 - \mathrm{WER} = \frac{C - I}{N} \tag{2.26}$$

als Qualitätsmaß benutzt. Bei genauer Betrachtung stellt man allerdings fest, dass beide Maße versuchen, die Gesamtfunktionalität des Erkenners in einer Zahl zu erfassen. Dabei gehen auch die Insertions in die entsprechende Rate mit ein, was nach Meinung des Autors ungünstig ist. Insertion-Test und Erkennungstest stellen wie oben beschrieben zwei unterschiedliche Aufgaben für den Erkenner dar. Ist bspw. die IR eines Erkenners hoch, würde die WER stark schwanken, je nachdem, ob das Korpus besonders viele bzw. wenige OOV-Wörter beinhaltet. Insbesondere kann der Autor sich nicht mit dem Weglassen von N_O im Nenner von (2.25) bzw. (2.26) einverstanden erklären, daraus entsteht auch die absurde Möglichkeit einer negativen WRR, bzw. einer WER über 100 %. Es ist, wie bereits erwähnt, zusätzlich eine Definitionsfrage, ob Deletions als Fehler zu interpretieren sind. Bezieht man N_O ein, wäre eine sinnvolle Messung mit

$$\mathrm{WER} = \frac{S + I}{N + N_O} \quad \text{oder} \quad \mathrm{WER} = \frac{S + D + I}{N + N_O} \tag{2.27}$$

bzw. die Worterkennungsrate

$$\mathrm{WRR} = \frac{C}{N + N_O} \quad \text{oder} \quad \mathrm{WRR} = \frac{C + C_{O,Rej}}{N + N_O} \tag{2.28}$$

anzusetzen.

2.3.6 Evaluations-Set-Up für diese Arbeit

Die vorliegenden Arbeit untersucht das Verhalten und die Verbesserung der Erkennungsleistung unter gestörten Umgebungsbedingungen. Dabei geht es zunächst darum, Erkennungsleistungen unverfälscht zu messen, bei verschiedenen Umgebungsbedingungen zu vergleichen und eine Verbesserung bei robustheitssteigernden Maßnahmen zu beobachten. Für diese Zwecke ist es günstig, die Roherkennungsrate nach Gleichung (2.19) bei ausgeschalteter Rückweisung anzuwenden. Ohne Rückweisung wird die reale Leistung des Klassifikators gemessen. Die Rückweisung ist für praktische Zwecke vorgesehen. Sie kann durch eine unterschiedliche Parametrierung die RR beeinflussen, indem sie als C klassifizierte Wörter mehr oder weniger zurückweist. Dadurch werden die Ergebnisse der Einzelversuche schwer vergleichbar. Ohne Rückweisung entstehen diese Probleme nicht, allerdings steigt dadurch die Fehlerrate. Das ist für die hier durchgeführten Experimente jedoch nicht von Bedeutung, denn es werden nicht die absoluten Werte, sondern allein die relativen Veränderungen von RR zwischen den Einzelexperimenten zur Bewertung in Betracht gezogen.

Für den durchzuführenden Erkennungstest wird ein Evaluationskorpus benutzt, welches sich aus einer Untermenge des APOLLO-Korpus[12] zusammensetzt. Es wurden zunächst 20 Sprecher ausgewählt, für die die Erkennung besonders gut funktioniert, um nicht bereits durch die ungünstige Aussprache o. ä. Fehler in die Evaluation einzubringen. Weiterhin wurden jeweils nur das Nahbesprechungsmikrofon und die Störbedingung *Stille* benutzt. Es entsteht ein Korpus von $17 \cdot 3 \cdot 20 = 1020$ Realisierungen.

Die OOV-Rückweisungsrate RR_O bzw. die Insertion Rate IR werden im Rahmen dieser Arbeit nicht gemessen. Die Untersuchung von Methoden der Rückweisung ist ein eigenes Forschungsthema.

[12] Die APOLLO-Evaluation [MHK+03] war ein kommerziell durchgeführter Vergleich von Spracherkennern aus dem Jahr 2003. Hintergrund war die geplante Herstellung einer Dunstabzugshaube mit Sprachbedienung durch ein namhaftes deutsches Unternehmen. An der Evaluation nahmen zehn Spracherkenner von verschiedenen Einrichtungen teil. Die Evaluation wurde unter realistischen Umgebungsbedingungen durchgeführt. Dafür wurde ein Korpus mit 8 simultan aufnehmenden, verschieden im Raum positionierten Mikrofonen eingespielt. Diese Prozedur wurde für 5 verschiedene Küchengeräuschszenarien sowie für 4 verschiedene Sprecherpositionen wiederholt. Der Wortschatz besteht aus 17 Kommandophrasen, z. B. *Apollo-Licht-Ein*. Es wurden 45 Sprecher unterschiedlichen Alters und unterschiedlicher Herkunft aufgenommen, die unter sämtlichen Störbedingungen und Positionen die 17 Kommandos jeweils 3 mal wiederholten. Insgesamt entstehen demnach $17 \cdot 3 \cdot 4 \cdot 5 = 1020$ Realisierungen pro Person. Bezieht man alle 45 Sprecher sowie die 8 Mikrofone mit ein, entsteht ein Gesamtkorpus von $1020 \cdot 8 \cdot 45 = 367200$ Realisierungen. Die Evaluation ergab im Übrigen, dass keiner der teilnehmenden Erkenner den praktischen Anforderungen in ausreichendem Maße gewachsen war.

3 Raumakustische Umgebungsbedingungen

3.1 Überblick

In diesem Kapitel wird die raumakustische Umgebung des Erkenners beschrieben. Dabei werden zunächst kurz einige hier benötigte physikalische Grundlagen der Akustik sowie Raumakustik wiederholt. Anschließend erfolgt eine systemtheoretische Betrachtung des Raumes, wobei die Raumimpulsantwort als wichtige systemtheoretische Beschreibungsmöglichkeit eingeführt wird. Methoden zur Messung von Raumimpulsantworten werden kurz beschrieben. Im Anschluss folgt der Überblick über eine statistische Studie zu raumakustischen Umgebungsbedingungen in Wohn- und Büroräumen. Als Ergebnis werden die Freiheitsgrade der raumakustischen Größen für die späteren Spracherkennungsexperimente eingegrenzt.

3.2 Sprecher-Raum-Mikrofon-Strecke

In diesem Abschnitt werden die wichtigsten akustischen Grundbeziehungen vorgestellt. Dabei beziehen sich die Ausführungen vorrangig auf [Kut00, Kut04]. Für die genauen Ableitungen der einzelnen Gleichungen wird auf die Standardliteratur verwiesen. Hier sollen primär die akustischen Eigenschaften von Räumen abgeleitet und mit den akustischen Beziehungen begründet werden. Ferner soll dargestellt werden, dass der Raum ein lineares System ist, zumindest näherungsweise. Die angestellten Betrachtungen gelten für Normalbedingungen in Gasen (vgl. DIN 1343 [DIN03a]). Weichen die Bedingungen in dem geringen Rahmen ab, der im Büro- und Wohnumfeld möglich ist, entstehen keine nennenswerten Auswirkungen.

3.2.1 Schallfeld und akustische Größen

Akustik ist die Wissenschaft des Schalls bzw. der Schallausbreitung. Schall stellt eine mechanische Schwingung in gasförmigen, flüssigen oder festen Medien dar, die hier jedoch auf den speziellen Fall der Luft eingegrenzt werden sollen.

In einem idealen Gas bilden die Größen Druck p (in Pascal, 1 Pa $= 1\ \frac{\mathrm{N}}{\mathrm{m}^2}$), Temperatur

T (in Kelvin K) sowie die Dichte ϱ (in $\frac{\mathrm{kg}}{\mathrm{m}^3}$) den Zustand des Gases ab. In der Form

$$\frac{p}{\varrho \cdot T} = R_s = \text{const.} \tag{3.1}$$

der allgemeinen Zustandsgleichung von Gasen wird durch das Verhältnis der 3 Zustandsgrößen die stoffspezifische Gaskonstante R_s ($R_{s,\mathrm{Luft}} = 287,06\ \frac{\mathrm{J}}{\mathrm{kg \cdot K}}$, für trockene Luft) bestimmt. Es ist zu erkennen, dass die Änderung einer der drei Zustandsgrößen auch zur Änderung mindestens einer der beiden anderen Größen führt, um das Verhältnis konstant zu belassen. Wird das Medium von einer Schallquelle oder -welle zum Schwingen angeregt, ändern sich die 3 Zustandsgrößen entsprechend; [Kut04] bezeichnet sie deshalb auch als Schallfeldgrößen[1]. Die Schwingung ist den Ruhegrößen überlagert[2]

$$p_{\mathrm{ges}}(t) = p_- + p_\sim(t) \tag{3.2}$$
$$\varrho_{\mathrm{ges}}(t) = \varrho_- + \varrho_\sim(t) \tag{3.3}$$
$$T_{\mathrm{ges}}(t) = T_- + T_\sim(t). \tag{3.4}$$

Eine eingebrachte Massebewegung führt für den Moment zu einer Druckschwankung von hohem zu niedrigem Druck, welche ihrerseits wieder zu einer Massebewegung führt. Beide Prozesse wechseln sich ab, sie schwingen. Die Schwingung des Drucks (Wechseldruck) $p_\sim(t)$ wird als Schalldruck und die Geschwindigkeit der Masseauslenkung wird als Schallschnelle $\vec{v}_\sim(t)$ bezeichnet. Für die meisten der folgenden Betrachtungen werden nur die Wechselgrößen herangezogen; die Kennzeichnung durch den Index $\sim$ entfällt aus Vereinfachungsgründen. Vereinfachend werden weiterhin die vektoriellen Größen (herrührend aus $\vec{v}$) als Skalare geschrieben (in ebenen und kugelförmigen Wellen möglich, da die Richtung der Vektoren stets die Ausbreitungsrichtung ist). Die Schwingung ist energetisch betrachtet das wechselseitige Überführen von potentieller (Druck) in kinetische Energie (Massebewegung). Der Vorgang kann durch ein Feder-Masse-System modelliert werden. Er breitet sich durch die Kopplungskräfte im Medium als Welle mit der Ausbreitungsgeschwindigkeit (oder Schallgeschwindigkeit)

$$c = \sqrt{\frac{\kappa \cdot p_{\mathrm{ges}}}{\varrho_{\mathrm{ges}}}} \approx \sqrt{\frac{\kappa \cdot p_-}{\varrho_-}} \tag{3.5}$$

aus und wird durch die Wellengleichung der Akustik beschrieben (Adiabatenexponent κ; $\kappa_{\mathrm{Luft}} = 1,4 = \text{const.}$). Die Schallgeschwindigkeit ist von der Temperatur abhängig

[1] Hier sind die Größen des Gaszustandes gemeint. In der Literatur findet man den Begriff Schallfeldgrößen auch als Oberbegriff für die akustischen Größen wie Schalldruck, Schallschnelle etc., von denen einige im Folgenden kurz vorgestellt werden.

[2] Die Änderung der Temperatur als Schwingung $T_\sim(t)$ ist nur der Vollständigkeit halber angegeben. Sie ist zwar durch die Druck- und Dichteschwankung physikalisch vorhanden [Kut04], spielt aber für weitere Betrachtungen keine Rolle. Wenn im Folgenden von Temperaturschwankungen die Rede ist, dann ist damit die statische Temperatur T_- gemeint.

(vgl. Gleichungen (3.1) und (3.5)); diese wird aber als stationär angenommen, d. h., während eines Schallereignisses, etwa der Äußerung eines Sprechers, ändert sie sich nicht. Die Näherungsformel

$$c = 331\frac{\mathrm{m}}{\mathrm{s}} + 0,6\frac{\mathrm{m}}{\mathrm{s}} \cdot \frac{\vartheta}{^\circ\mathrm{C}} \tag{3.6}$$

gilt innerhalb der üblichen Umgebungsbedingungen ($-20^\circ\mathrm{C} < \vartheta < 40^\circ\mathrm{C}$). Weiterhin wird die Schallgeschwindigkeit hier als frequenzunabhängig betrachtet, da Luft ein nicht dispersives Medium ist [Kut04]. Der Zusammenhang zwischen p und v wird durch die mediumspezifische Schallimpedanz Z (oder Wellenwiderstand) ausgedrückt und ergibt sich zu

$$Z = \frac{p}{v} = \varrho_- \cdot c \approx \sqrt{\kappa \cdot p_- \cdot \varrho_-}. \tag{3.7}$$

Die sich ausbreitende Welle transportiert die Energie, die eine Schallquelle ursprünglich in das Medium abgegeben hat. Dabei wird die Menge an Energie, die sich an einem Raumpunkt befindet, bezogen auf sein Volumen als Energiedichte w bezeichnet. Sie wird mit der Schallgeschwindigkeit von der Welle transportiert. In

$$w = \frac{I}{c} \tag{3.8}$$

beschreibt die Schallintensität I (oder auch Energieflussdichte) die flächenbezogene Menge an Energie pro Zeiteinheit, die durch eine Oberfläche normal zur Ausbreitungsrichtung transportiert wird. Demnach entspricht die Intensität einer flächenbezogenen Leistung, die für jeden Punkt des Schallfeldes nach

$$I = \overline{vp} = \frac{\tilde{p}^2}{Z} \tag{3.9}$$

berechnet werden kann. $\tilde{p}$ ist dabei der Effektivwert des Schalldrucks und $\overline{vp}$ der zeitliche Mittelwert des Produktes der beiden Schallfeldgrößen.

Einfach messbar (durch Mikrofone) ist meist der Schalldruck, der deshalb auch die wichtigste Größe in der Akustik ist. Die Beschreibung des Verhaltens der Schallfelder geschieht jedoch energetisch anhand der Energiedichte. Es ergibt sich der proportionale Zusammenhang zwischen Effektivwertquadrat des Schalldruckes und der Energiedichte nach (3.8) und (3.9) zu

$$\tilde{p}^2 = Z \cdot c \cdot w \tag{3.10}$$

$$\tilde{p}^2 \sim w. \tag{3.11}$$

Die folgenden Gleichungen beziehen sich deshalb besonders auf die Ausbreitung des Schalldrucks und nicht anderer akustischen Größen. Da der Schalldruck in seinen

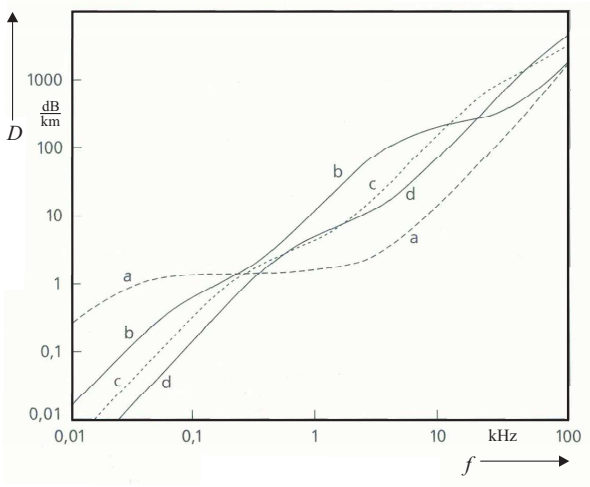

Abbildung 3.1 – Dämpfungsmaß von Luft in dB/km bei 20 °C und Normaldruck. Parameter: relative Luftfeuchtigkeit (a) 0 %, (b) 10 %, (c) 40 % und (d) 100 %. Graphik aus [Kut04].

hörbaren Grenzen einen großen Dynamikbereich überspannt, ist die Angabe als Pegel, Schalldruckpegel L_p, üblich. Dabei wird der Schalldruck auf die international genormte Hörschwelle $p_0 = \tilde{p}_{H_0} = 2 \cdot 10^{-5}$Pa bezogen[3]

$$L_p = 20 \lg \frac{\tilde{p}}{p_0} \text{dB}. \tag{3.12}$$

Die Lösung der Wellengleichung für den Schalldruck für eine einzelne harmonische Basisschwingung einer ebenen Welle (angenommene Ausbreitung in x-Richtung) lautet

$$p(x,t) = \hat{p} \cos\left[\omega\left(\frac{x}{c} - t\right)\right]. \tag{3.13}$$

Es ist zu erkennen, dass die Funktion für einen festen Punkt x_i nur die Zeitabhängigkeit besitzt. Durch den feststehenden Ort x ergibt sich in (3.13) eine frequenzabhängige Phasenverschiebung. Die Laufzeitverzögerung ist frequenzunabhängig. Dies gilt, solange $c(f) = \text{const.}$ ist, wovon, wie oben beschrieben, ausgegangen werden kann. Das Frequenzgemisch eines Sprachsignales (oder eines anderen auditiven Signales) wird somit für die Schwingungen aller Frequenzen (additive Überlagerung) vom Medium gleichförmig übertragen.

[3] In der Standardliteratur (z. B. [MM04, Kut04]) wird p_0 als Bezugsgröße für einen (Schall-)Druck angegeben, jedoch nicht als Effektivwert. Deshalb wird hier für den Effektivwert der Hörschwelle zur Unterscheidung $\tilde{p}_{H_0}$ angegeben.

Tabelle 3.1 – Relevanz der Luftdämpfung bei verschiedenen Distanzen. Die Werte wurden nach (3.16) mit abgelesenen Werten aus Abbildung 3.1 (Kurve (b)) berechnet.

f	200 Hz	8000 Hz
$D(f)$	1 $\frac{\text{dB}}{\text{km}}$	100 $\frac{\text{dB}}{\text{km}}$
x	$\Delta L_p(x, 200\text{ Hz})$	$\Delta L_p(x, 8000\text{ Hz})$
10 m	0,01 dB	1 dB
100 m	0,1 dB	10 dB
200 m	0,2 dB	20 dB
500 m	0,5 dB	50 dB

In (3.13) wird die Ausbreitung der Welle als verlustfreier Energietransport betrachtet. Diese Annahme ist noch etwas ungenau. Aufgrund von sogenannten dissipativen Prozessen wird ständig Wirkleistung vom Medium in Wärme umgesetzt, es wird Energie absorbiert. Diese Absorption ist von verschiedenen Eigenschaften abhängig, wobei die Luftfeuchtigkeit eine besonders große Bedeutung hat. Die Absorption ist frequenzabhängig und kann mit dem Dämpfungsfaktor $d(x, f)$ beschrieben werden

$$d(x, f) = e^{-\frac{m(f)}{2}x}. \tag{3.14}$$

Er beschreibt, dass die Dämpfung der Welle mit größer werdender Strecke x ansteigt, da sie mehr und mehr Medium durchläuft. Die Größe $m(f)$ wird in der Literatur als mediumspezifische Dämpfungskonstante bezeichnet. Sie ist ebenfalls frequenzabhängig. Gleichung (3.13) wird somit zu

$$p(x, t) = \hat{p}\ e^{-\frac{m(f)}{2}x}\ \cos[\omega(\frac{x}{c} - t)]. \tag{3.15}$$

Bestimmt man die Pegelabnahme bezogen auf die zurückgelegte Strecke, beschreibt

$$\Delta L_p(x, f) = D(f) \cdot x \tag{3.16}$$

den gewünschten Zusammenhang durch Pegelbildung von (3.14). $D(f)$ wird als Dämpfungsmaß bezeichnet und ergibt sich nach

$$D(f) = 10\ m(f)\ \lg e\ \frac{\text{dB}}{\text{m}}. \tag{3.17}$$

Es beschreibt die Pegelabnahme bezogen auf eine Strecke. Abbildung 3.1 zeigt, dass die Dämpfung für trockene Luft (Kurve (a)) im hier relevanten Frequenzbereich ($f < \frac{f_s}{2} = 8$ kHz) fast konstant ist, jedoch bei geringer Luftfeuchte von 10 % bereits eine sehr starke Frequenzcharakteristik hat. In Tabelle 3.1 wird beispielhaft dargestellt, dass in normaler Kommunikationsentfernung von wenigen Metern die Luftdämpfung

bedeutungslos ist, jedoch in einigen hundert Metern keinesfalls mehr vernachlässigt werden kann. Arbeiten gehen hier u. a. auf [BBE72] zurück, wobei [Kut00] die neuere Literaturstelle [BSZ+95] zitiert. Eine frequenzunabhängige Dämpfung wäre unproblematisch. Durch die Frequenzabhängigkeit der Dämpfung wird das Signal verzerrt, was in größeren Distanzen zu einer Überbetonung der Tiefen bzw. Dämpfung der Höhen führt.

3.2.2 Schallfeld in Räumen

Nachdem wichtige Größen allgemein eingeführt wurden, werden sie nun speziell auf die Schallausbreitung in Räumen angewandt. Dabei wird zwischen dem direkten und dem diffusen Schallfeld unterschieden.

Direktschallfeld Das direkte Schallfeld beschreibt den Schall, der vom Medium ohne Hindernis direkt von einer abstrahlenden Quelle zum Empfänger transportiert wird. In Gleichung (3.13) ist $\hat{p}$ nicht von x abhängig, da es sich um eine ebene Welle handelt. Der Mensch erzeugt als Quelle jedoch keine ebene Welle, sondern soll zunächst als Quelle einer kugelförmigen Welle angenähert werden. Die Quellfunktion einer kugelförmigen Welle kann durch die Volumengeschwindigkeitsfunktion $Q(t)$ einer gedachten Punktquelle beschrieben werden, die ein gedachtes Volumen ein- bzw. ausatmet. Die Wellengleichung erhält damit laut [Kut00] für den Schalldruck eine Lösung der Form (Ausbreitungsrichtung radial, $\vec{r}$ vereinfacht als Abstand $r = |\vec{r}|$)

$$p(r,t) = \frac{\varrho_-}{4\pi r}\dot{Q}\left[\omega\left(\frac{r}{c}-t\right)\right]. \tag{3.18}$$

Die Ableitung $\dot{Q}(t)$ kann dabei als eine dem Schalldruck proportionale Größe gesehen werden. Für eine einzelne harmonische Schwingung ergibt sich

$$Q(t) = \hat{Q}\cos(\omega t) \tag{3.19}$$

und demnach

$$p(r,t) = \frac{\varrho_- \cdot \omega\hat{Q}}{4\pi r}\sin\left[\omega\left(\frac{r}{c}-t\right)\right] \tag{3.20}$$

$$= \hat{p}(r)\sin\left[\omega\left(\frac{r}{c}-t\right)\right]. \tag{3.21}$$

Die Dämpfung der Luft soll im Direktschallfeld vernachlässigt werden, da die Distanzen in Räumen im Wohn- und Büroumfeld normalerweise geringer als 10 m sind. Der Unterschied zur ebenen Welle ist die Entfernungsabhängigkeit $\hat{p}(r)$ in (3.21). Benutzt man den Effektivwert $\tilde{p}(r)$, so kann die Intensität an einem beliebigen Punkt nach (3.9)

berechnet werden. Die abgestrahlte Schallleistung der Quelle P_Q lässt sich bestimmen, wenn die Intensität über eine Kugeloberfläche $4\pi r^2$ um die Quelle herum integriert wird

$$P_\mathrm{Q} = 4\pi r^2 \cdot I(r) = 4\pi r^2 \cdot \frac{\tilde{p}^2}{Z}. \tag{3.22}$$

Substituiert man die Intensität mit (3.8), entsteht nach Umstellen von (3.22) das Abstandsgesetz für die Energiedichte des direkten Schallfeldes (Index D)

$$w_\mathrm{D}(r) = \frac{P_\mathrm{Q}}{4\pi c r^2} \tag{3.23}$$

$$w_\mathrm{D}(r) \sim \frac{1}{r^2}. \tag{3.24}$$

Für den Schalldruck lässt sich mit (3.22) oder (3.11) ebenfalls ein Abstandsgesetz ableiten

$$\tilde{p}_\mathrm{D}(r) = \frac{1}{2r}\sqrt{\frac{\varrho_- \cdot c \cdot P_\mathrm{Q}}{\pi}} \tag{3.25}$$

$$= \frac{1}{2r}\sqrt{\frac{Z \cdot P_\mathrm{Q}}{\pi}} \tag{3.26}$$

$$\tilde{p}_\mathrm{D}(r) \sim \frac{1}{r}. \tag{3.27}$$

Aus dem Abstandsgesetz kann

$$\frac{\tilde{p}_1}{\tilde{p}_2} = \frac{r_2}{r_1} \tag{3.28}$$

geschlussfolgert werden. Mit dem Schalldruckpegel ergibt sich (3.28) zu

$$L_{p_2} = L_{p_1} - 20\lg\frac{r_2}{r_1}\mathrm{dB}. \tag{3.29}$$

Die getroffenen Annahmen befassen sich mit einer idealisierten Schallausbreitung. Diese findet so nur im sogenannten Freifeld statt, wobei sich die Schallwellen bis ins Unendliche ungestört ausbreiten. Abgesehen von ständig vorhandenen (Stör-)Geräuschen stellt auch eine Freiluftfläche nur bedingt ein idealisiertes Freifeld dar, da Schallwellen vom Boden reflektiert werden. In gewissen Grenzen lässt sich ein solches Freifeld in einem reflexionsfreien Raum simulieren.

Diffuses Schallfeld Innerhalb von geschlossenen Räumen ist die ungestörte Wellenausbreitung räumlich begrenzt. Beim Auftreffen einer Welle auf Hindernisse wird nur ein Teil der Welle reflektiert, der andere wird absorbiert[4]. Die Energie einer reflektierten Welle beträgt nur noch $\rho = 1 - \alpha$ mal der Energie der zugehörigen Originalwelle. ρ ist der Reflexionsgrad. Der Absorptionsgrad α ist abhängig von der Frequenz des Schalles, von der Oberfläche und vom Material des Hindernisses. Tabelle 3.2 zeigt beispielhaft die Frequenzabhängigkeit der Absorptionsgrade einiger typischer Materialien. Das Produkt einer Teilfläche S_i und ihres Absorptionsgrades α_i ergibt ihre äquivalente Absorptionsfläche A_i. Die Summe über alle Teilflächen eines Raumes ergibt die äquivalente Absorptionsfläche A des Raumes

$$A(f) = \sum_i S_i \cdot \alpha_i(f), \tag{3.30}$$

die dessen Absorptionsvermögen charakterisiert. Genau genommen ist A wegen $\alpha(f) \neq$ const. noch frequenzabhängig, was in der Literatur meist nicht explizit erwähnt wird.

Durch zahlreiche (möglichst gleichverteilte) Reflexionen entsteht ein diffuses Schallfeld, der Hall. Bei idealer Diffusität ist die Energiedichte des Schallfeldes an jedem Raumpunkt gleich

$$w_\mathrm{R}(\vec{r}) = \text{const.} \tag{3.31}$$

Der Index R verweist auf das Wort Reverberation (englisch für Hall). Die Bedingung (3.31) ist in der Praxis nicht immer so ideal gegeben. Allerdings kann sie meist, zumindest näherungsweise angenommen werden [Kut00]. Nur mit der Annahme eines konstanten Diffusfeldes können die Gesetze der Raumakustik in einfacher Weise formuliert werden [Kut00], deshalb soll diese Idealisierung auch für den Rahmen dieser Arbeit angenommen werden. Die Frequenzcharakteristik (das Leistungsdichtespektrum) des diffusen Hallfeldes muss nicht konstant sein, sie folgt zum einen dem Quellsignal, wird jedoch durch die Frequenzabhängigkeit der Reflexionen verfärbt ($A(f) \neq$ const.). Ein zweiter Grund für die Verfärbung ist die Frequenzabhängigkeit der Luftdämpfung bei größeren Distanzen, da die einzelnen Diffuskomponenten nach ihrer Abstrahlung durch das Quellsignal über mehrere Reflexionspunkte bereits einige hundert Meter zurückgelegt haben können (vgl. Tabelle 3.1).

Nach einigen Definitionen leitet [Kut00] die Differenzialgleichung

$$P_\mathrm{Q}(t) = V\frac{\mathrm{d}w_\mathrm{R}}{\mathrm{d}t} + \frac{cA}{4}w_\mathrm{R} \tag{3.32}$$

zur Aufnahme der von einer Quelle erzeugten Schallleistung P_Q vom Raum ab (vgl. oben: Interpretation der Energiedichte pro Zeiteinheit als flächenbezogene Leistung →

[4] Es ist anzumerken, dass hier die Absorption im Sinne des Verlustes von nicht reflektiertem Schall aus Sicht des Raumes gemeint ist. In der Akustik spricht man bei einem Hindernis, wie einer Wand, von den drei Komponenten Reflexion, Absorption und Transmission. Die beiden Letzteren werden hier zusammengefasst.

Tabelle 3.2 – Absorptionsgrade verschiedener Wandmaterialien. Die Messungen wurden von der Physikalischen Technischen Anstalt Braunschweig durchgeführt. Absorptionsgrade über 1 sind durch Unvollkommenheiten des Messverfahrens bedingt, 1Rayl = $10\frac{\mathrm{Ns}}{\mathrm{m}^3}$. Traditionelle Massivbauten tendieren zur größeren Absorption mit steigender Frequenz. Bei modernen Leichtbauten werden tiefe Frequenzen an Holz-, Gipskarton- oder Glaswänden weniger reflektiert als hohe (Tabelle und Bemerkung aus [HM94a]).

Material	Absorptionsgrad bei					
	125	250	500	1000	2000	4000 Hz
Harte Flächen (Putz, Mauerwerk, harte Fußböden)[b]	0,02	0,02	0,03	0,03	0,04	0,04
Teppich in Schlingenwebart, 4,5 mm dick, imprägniert, direkt auf Boden[a]	–	0,02	0,04	0,15	0,36	0,32
Satinvorhang 82 Rayl, 20 cm vor Wand, 1,5fache Faltung[a]	0,09	0,55	1,03[d]	0,89	0,93	0,92
gebundene Mineralfaserplatte 12 Rayl/cm, 30 mm dick[a]	–	0,44	0,84	0,84	0,93	0,88
wie oben, aber mit 50 mm lichtem Wandabstand montiert[a]	–	0,73	1,00	0,89	0,82	0,84
Mineralfaserplatte 20 Rayl/cm, 50 mm dick, mit 50 mm lichtem Wandabstand, sichtseitig mit 6 mm Sperrholz abgedeckt[a]	0,40	0,53	0,29	0,18	0,11	0,11
Holz, 1,6 cm dick, auf 4 cm Holzlatten[b]	0,18	0,12	0,10	0,09	0,08	0,07
18 mm Gipskartonplatte, 16 kg/m², 400 mm vor starrer Wand[c]	0,10	0,09	0,05	0,05	0,07	0,04
wie vor, hinterlegt mit 30 mm Mineralfasermatte 1,05 kg/m², 7,5 Rayl/cm[c]	0,18	0,10	0,08	0,07	0,10	0,10
geschlossenes Doppelfenster	0,10	0,04	0,03	0,02	0,02	0,02
Gipskartonlochplatte 19,6% Lochflächenanteil, 15 mm Lochdurchmesser, 100 mm vor starrer Wand, hinterlegt mit Faservlies 12 Rayl, Mineralfasermatte 1,05 kg/m², 7,5 Rayl[a]	0,30	0,69	1,01[d]	0,81	0,66	0,62
Metallpaneele aus 0,5 mm Alublech, 85 mm breit, freie Schlitzbreite zwischen Paneelen 15 mm, 164 mm Abstand vor starrer Wand, hinterlegt mit 20 mm Mineralfaserplatte 2,5 kg/m², 60 Rayl[a]	0,25	0,59	0,81	0,64	0,26	0,17

Intensität). Es entsteht ein stabiler Zustand, wenn die diffuse Energiedichte w_R mit der Zeit nicht mehr ansteigt ($\frac{\mathrm{d}w_\mathrm{R}}{\mathrm{d}t} = 0$). Dann existiert das Gleichgewicht zwischen erzeugter Quellleistung und der in der äquivalenten Absorptionsfläche verschwindenden Leistung. Daraus kann

$$w_\mathrm{R} = \frac{4P_\mathrm{Q}}{cA} \tag{3.33}$$

berechnet werden. Bei Abschalten der Quelle ($P_\mathrm{Q} = 0$) zum Zeitpunkt t_0 wird (3.32) homogen und hat die Lösung

$$w_\mathrm{R}(t) = w_\mathrm{R}(t_0) \cdot e^{-\frac{cA}{4V}t}. \tag{3.34}$$

Die Zeit, in der $w_\mathrm{R}(t)$ nach Abschalten einer Schallquelle auf seinen millionsten Teil gefallen ist, wird als Nachhallzeit T_{60} bezeichnet. Der Index 60 soll darauf hindeuten, dass es sich dabei um den Abfall von 60 dB handelt[5]. Setzt man den 60-dB-Abfall bei T_{60} in (3.34) ein

$$10^{-6} = e^{-\frac{cA}{4V}T_{60}} \tag{3.35}$$

ergibt sich nach Umstellen

$$T_{60} = 6 \ln 10 \frac{4}{c}\frac{V}{A} = 0,163 \frac{V}{A}\ \frac{\mathrm{s}}{\mathrm{m}}. \tag{3.36}$$

Die Nachhallzeit ist die wichtigste Größe in der Raumakustik und wurde von C. W. Sabine Ende des 19. Jahrhunderts eingeführt ((3.36) ist die nach ihm benannte Formel). Da, wie erwähnt, $A(f)$ frequenzabhängig ist, ergibt sich auch eine Frequenzabhängigkeit der Nachhallzeit ($T_{60}(f) \neq$ const.). Durch Substitution von A in (3.33) mit (3.36) erhält man

$$w_\mathrm{R} = \frac{T_{60}}{13,8 \cdot V} P_\mathrm{Q}. \tag{3.37}$$

Mit (3.36) wird (3.34) zu

$$w_\mathrm{R}(t) = w_\mathrm{R}(t_0) \cdot e^{-\frac{6 \ln 10}{T_{60}} t} \tag{3.38}$$

$$= w_\mathrm{R}(t_0) \cdot e^{-\frac{13,8}{T_{60}} t}. \tag{3.39}$$

Stellt man w_R als Pegelgröße in dB dar, ergibt sich durch die Logarithmierung von (3.34) eine fallende Gerade

$$L_{w_\mathrm{R}}(t) = L_{w_\mathrm{R}}(t_0) - 10 \lg e \cdot \frac{cA}{4V} t\ \ \mathrm{dB} \tag{3.40}$$

mit dem Anstieg

$$\frac{-60\ \mathrm{dB}}{T_{60}} = -10 \lg e \cdot \frac{cA}{4V}\ \ \mathrm{dB}. \tag{3.41}$$

Nach Umstellen entsteht wieder die Sabine'sche Formel

$$T_{60} = \frac{6}{\lg e}\frac{4}{c}\frac{V}{A} = 0,163 \frac{V}{A}\ \ \frac{\mathrm{s}}{\mathrm{m}}. \tag{3.42}$$

Aufgrund der fallenden Gerade gibt es verschiedene Verfahren zur Messung der Nachhallzeit nach Abschalten einer Schallquelle, die sich mit ihrem Anstieg befassen (u. a. auch das später beschriebene Schröder-Integral, vgl. Abschnitt 3.3.5).

[5] Oft wird auch nur die Notation T verwendet. In der englischsprachigen Literatur findet man auch RT für Reverberation-Time. Manchmal findet man auch die Notation T_{20} oder T_{30}, häufig bei akustischer Messtechnik. Sie deutet an, dass der Abfall nur über 20 bzw. 30 dB beobachtet und die gemessene Zeit danach mit 3 bzw. 2 multipliziert wurde. Diese Notation ist etwas ungünstig, da sie mit der ursprünglichen Motivation des Index 60 leicht zu verwechseln ist und deshalb missverstanden werden kann.

Gesamtschallfeld Die Energiedichte des gesamten Schallfeldes w ergibt sich aus der Superposition von diffuser und direkter Schallenergiedichte und mit entsprechendem Einsetzen von (3.24) und (3.33) zu

$$w(r) = w_{\mathrm{D}}(r) + w_{\mathrm{R}} \tag{3.43}$$

$$w(r) = \frac{P_{\mathrm{Q}}}{c}\left(\frac{1}{4\pi r^2} + \frac{4}{A}\right), \tag{3.44}$$

wobei $w(r)$ wieder vom Radius abhängt (kugelförmig), herrührend von $w_{\mathrm{D}}(r)$. Bei mehreren Quellen addieren sich die entsprechenden Felder der Einzelquellen. Bildet man das Verhältnis aus direkter und diffuser Schallenergiedichte, so ergibt sich ein Direktschall-Hall-Abstand (engl.: Direct-to-Reverberation Ratio – DRR, zur Begrifflichkeit DRR siehe auch Abschnitt 4.4.11)

$$\mathrm{DRR} = 10\lg\frac{w_{\mathrm{D}}}{w_{\mathrm{R}}}\ \mathrm{dB} \tag{3.45}$$

$$= 10\lg\frac{A}{16\pi r^2}\ \mathrm{dB} \tag{3.46}$$

$$= 10\lg\frac{0,163\frac{\mathrm{s}}{\mathrm{m}}\cdot V}{16\pi r^2 T_{60}}\ \mathrm{dB} \tag{3.47}$$

$$= (-24,9 + 10\lg\frac{V}{\mathrm{m}^3} - 20\lg\frac{r}{\mathrm{m}} - 10\lg\frac{T_{60}}{\mathrm{s}})\ \mathrm{dB}. \tag{3.48}$$

Gleichung (3.48) zeigt die Abhängigkeiten des DRR von den akustisch relevanten Größen. Der DRR selbst ist nicht von der Quellleistung P_Q abhängig. Bei Pegelschreibweise erscheint sie als additiver Wert in $L_{w_{\mathrm{R}}}$ und $L_{w_{\mathrm{D}}}$. Sie wird in (3.45) gekürzt. Abbildung 3.2 zeigt $w_{\mathrm{D}}(r)$ und w_{R} in Pegeldarstellung für ein bestimmtes P_{Q}. Ändert man P_{Q} (bspw. um +10 dB), so ändert sich an den Verhältnissen in der Graphik nichts, die Ordinate wird lediglich um die Änderung (+10 dB) verschoben. Bei gegebenem P_{Q} ist $w_{\mathrm{D}}(r)$ nicht vom Raum abhängig. Dagegen ist das abstands- bzw. ortsunabhängige w_{R} typisch für einen Raum, was hier beispielhaft für drei unterschiedlich große Räume, deren Nachhallzeit als gleich angenommen wird, dargestellt ist. Der Radius, an dem w_{R} und w_{D} gleich groß sind, ist als Hallradius r_{R} bekannt und ergibt sich aus Gleichsetzen von (3.24) und (3.33)

$$\frac{4P_{\mathrm{Q}}}{cA} = \frac{P_{\mathrm{Q}}}{4\pi c r_{\mathrm{R}}^2} \tag{3.49}$$

$$r_{\mathrm{R}} = \frac{1}{4}\sqrt{\frac{A}{\pi}}, \tag{3.50}$$

und durch die Substitution von A mit der Sabine'schen Formel

$$r_{\mathrm{R}} = \frac{1}{10}\sqrt{\frac{V}{\pi T_{60}}\frac{\mathrm{s}}{\mathrm{m}}} \approx \sqrt{0,0032\frac{V}{T_{60}}\ \frac{\mathrm{s}}{\mathrm{m}}}. \tag{3.51}$$

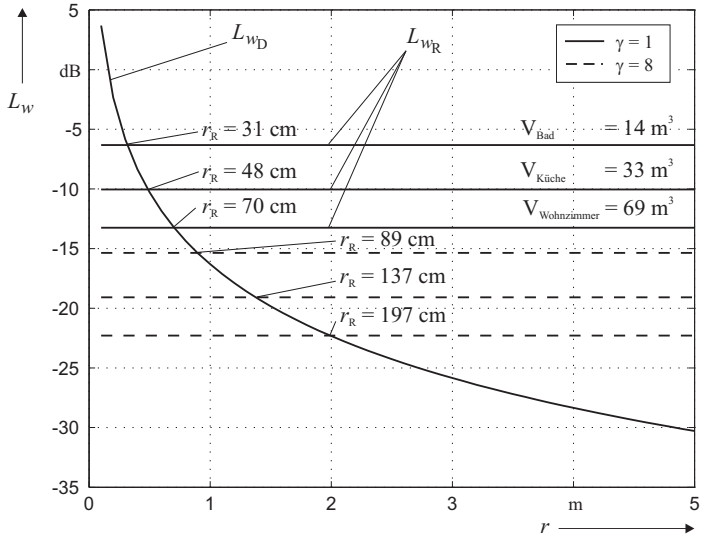

Abbildung 3.2 – Pegeldarstellung von Direktschall-Energiedichte L_{w_D} im Vergleich zur Diffusschall-Energiedichte L_{w_R}. Das Beispiel zeigt das simulierte Verhalten in Räumen aus der Studie in Tabelle 3.3. D. h., T_{60} = 0,43 s (Gesamtmittelwert), Mittelwerte der Volumina aus den Klassen Bad, Küche und Wohnzimmer (Werte vgl. Tabelle 3.3). Die Berechnung von L_{w_D} und L_{w_R} erfolgt jeweils durch Pegelberechnung der Gleichungen (3.24) bzw. (3.37). Die gestrichelten Linien zeigen den Unterschied einer ungerichteten Quelle oder Senke ($\gamma = 1$) im Vergleich zu einer Quelle oder Senke mit einer Richtcharakteristik von $\gamma = 8$.

Wenn die Quelle keine kugelförmige Richtcharakteristik hat, ist die Energie des Schalls in Hauptrichtung, also am Mikrofon, größer als der durchschnittlich in alle Richtungen abgestrahlte Schall. Eine ähnliche Interpretation entsteht bei einem Mikrofon mit Richtcharakteristik. Es nimmt den Schall aus der Hauptrichtung stärker wahr als den Schall aus den anderen Richtungen. Bei beiden Fällen sinkt die wahrgenommene w_R im Vergleich zu einer entsprechenden Kugelcharakteristik. Daraus folgt, dass der Schnittpunkt von w_D und w_R weiter von der Quelle wegrückt, d. h., der Hallradius vergrößert sich (Abbildung 3.2). Ein Maß für die Richtwirkung von Quelle oder Mikrofon ist der Bündelungsgrad γ (vgl. Abschnitt 3.2.3.1), der laut [Kut00] in die Berechnung des Hallradius einbezogen wird

$$r_R = \frac{1}{10}\sqrt{\frac{\gamma V}{\pi T_{60}}\,\frac{\mathrm{s}}{\mathrm{m}}}. \tag{3.52}$$

Diese γ-Abhängigkeit ist auch für den Fall des hier beschriebenen Anwendungsszenarios interessant, da, wie in den Abschnitten 3.2.3.2 und 3.2.3.3 gezeigt wird, Mensch und

Mikrofon keine Kugelcharakteristik besitzen. Abbildung 3.2 zeigt realistische Werte für den Hallradius im Wohn- und Bürobereich, da hier die Mittelwerte aus der Studie in Tabelle 3.3 verwendet wurden.

Auch der Hallradius ist frequenzabhängig. Das ergibt sich aus der Frequenzabhängigkeit von γ und T_{60}. Mit dem Hallradius (3.50) ergibt sich (3.24) zu

$$w_{\mathrm{D}}(r) = \frac{4P_{\mathrm{Q}}}{cA} \cdot \frac{r_{\mathrm{R}}^2}{r^2}, \tag{3.53}$$

die Gesamtenergiedichte zu

$$w(r) = \frac{4P_{\mathrm{Q}}}{cA}\left(1 + \frac{r_{\mathrm{R}}^2}{r^2}\right), \tag{3.54}$$

bzw. mit (3.10) der Schalldruck zu

$$\tilde{p}^2(r) = \varrho_- \frac{4P_{\mathrm{Q}}}{cA}\left(1 + \frac{r_{\mathrm{R}}^2}{r^2}\right) \tag{3.55}$$

$$\tilde{p}^2(r) \sim \left(1 + \frac{r_{\mathrm{R}}^2}{r^2}\right). \tag{3.56}$$

Mit (3.51) ergibt sich für die Berechnung des DRR in (3.47)

$$\mathrm{DRR} = 10 \lg \frac{r_{\mathrm{R}}^2}{r^2}\ \mathrm{dB} \tag{3.57}$$

und

$$\frac{w_{\mathrm{D}}}{w_{\mathrm{R}}} = \frac{r_{\mathrm{R}}^2}{r^2}. \tag{3.58}$$

Obwohl diese Beziehung sehr trivial erscheint, ist es nicht möglich, eine einfache Funktion für die Störung durch den Hall abzuleiten, die nur von r abhängig ist, da in r_{R}^2 laut (3.52) die variablen Werte Volumen, Nachhallzeit und Richtcharakteristik versteckt sind. Demnach wäre die wichtigste raumbeschreibende akustische Größe der Hallradius und nicht wie üblich die Nachhallzeit. Da man nur das Gesamtschallfeld messen kann, sind direktes und diffuses Schallfeld in Räumen separat nicht messbar. Abbildung 3.2 veranschaulicht allerdings, dass man in genügend großem Abstand von r_{R} näherungsweise

$$\begin{aligned} \tilde{p}_{\mathrm{D}}(r) &\approx \tilde{p}(r << r_{\mathrm{R}}) \\ \tilde{p}_{\mathrm{R}}(r) &\approx \tilde{p}(r >> r_{\mathrm{R}}) \end{aligned} \tag{3.59}$$

messen kann. Es soll an dieser Stelle noch bemerkt werden, dass alle Annahmen Modellvorstellungen sind und daher in der Praxis nicht immer exakt nachgewiesen werden können.

3.2.3 Eigenschaften von Quelle und Mikrofon

3.2.3.1 Richtcharakteristiken

Richtcharakteristiken entstehen bei der räumlichen Abstrahlung (Quelle) oder dem Empfang (Senke) von Energie. Die Prinzipien sind dabei für verschiedene Energieformen gleich, egal ob es sich um Lichtquellen, Schallquellen, Mikrofone, Antennen usw. handelt. Es wird eine räumliche Verteilung einer energie- oder intensitätsbeschreibenden Größe gemessen; bei der akustischen Richtcharakteristik ist es der Effektivwert des Schalldrucks. Dabei entsteht ein Richtdiagramm, in dem die Richtungsfunktion $\Gamma(\theta)$ (oder auch als Funktion des räumlichen Winkels $\Gamma(\theta, \phi)$) dargestellt wird (Darstellung meist als Richtungsmaß in dB)

$$\Gamma(\theta, \phi) = 10 \lg \frac{\tilde{p}(\theta, \phi)}{\tilde{p}_{0^\circ}} \bigg|_f \quad \text{dB}. \tag{3.60}$$

Die Richtungsfunktion ist meist frequenzabhängig, deshalb gilt das Richtdiagramm für eine Frequenz. Oft werden auch mehrere Graphen für verschiedene Frequenzen in einem Diagramm dargestellt, um die Richtcharakteristik vollständiger darzustellen. Eine weitere Größe, die die Richtcharakteristik beschreibt, ist der Bündelungsgrad $\gamma(f)$ (auch Bündelungsmaß in dB)

$$\gamma(f) = \frac{\tilde{p}_{0^\circ}(f)}{\frac{1}{I} \sum_{i=1}^{I} \tilde{p}_i(f)}. \tag{3.61}$$

Er ermöglicht es, eine bessere frequenzabhängige Vorstellung der Richtcharakteristik zu erhalten. Er wird aus dem Verhältnis des in Hauptstrahlrichtung gemessenen Schalldruckes $\tilde{p}_{0^\circ}(f)$ zum Mittelwert des Schalldruckes in alle Richtungen gemessen. Zur Berechnung des Mittelwertes wird $\tilde{p}(f)$ formal über eine Kugeloberfläche integriert und durch den Kugeloberflächeninhalt dividiert. In der Praxis ist es sinnvoll, das arithmetische Mittel von $\tilde{p}_i(f)$ einer geeigneten Anzahl I von gerasterten Messpunkten i zu bilden (Gleichung (3.61)).

3.2.3.2 Eigenschaften der Quelle

Im Spracherkennungsszenario ist der Mensch die Schallquelle. Er erzeugt Sprache.

Richtcharakteristik Im Hinblick auf die akustischen Eigenschaften der Sprache ist die Richtcharakteristik der menschlichen Schallabstrahlung von Bedeutung. Richtdiagramme für die menschliche Schallabstrahlung sind in Abbildung 3.3 dargestellt. Die Richtcharakteristik beim Menschen ist typisch für Wellenausbreitungsvorgänge. Nach

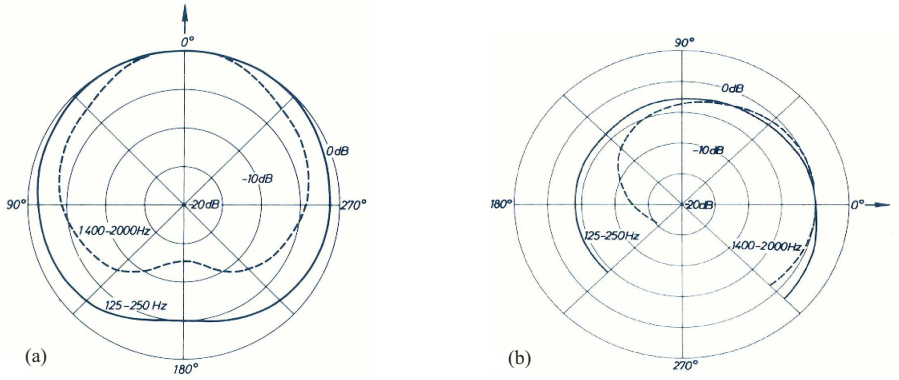

Abbildung 3.3 – Richtcharakteristik der menschlichen Schallabstrahlung. (a) horizontales Richtdiagramm. (b) vertikales Richtdiagramm. Graphiken aus [Kut00].

vorn werden alle Frequenzen gleich abgestrahlt, zur Seite und nach hinten werden hohe Frequenzen stärker gedämpft als tiefe. Dieses Verhalten wird bspw. zur Ermittelung der Kopforientierung bei akustischer Personenlokalisierung genutzt [ASNH07]. Zur menschlichen Richtcharakteristik existiert keine umfangreiche Literatur [HV04]. Es gibt nur vereinzelte Studien [Bri98, DF39, Fla60, McK86, SI73, MP78, HV04]. Allerdings findet man darin keinen Bündelungsgrad γ (vgl. (3.61)), wie man ihn in (3.52) benötigen würde. In Abbildung 3.3 (an anderen Literaturstellen ähnlich) ist das Richtdiagramm nur für Frequenzen bis 2000 Hz gegeben. Für höhere Frequenzen lässt sich aufgrund der Beugungseigenschaften von Wellen eine noch stärkere Richtcharakteristik vermuten.

Frequenzcharakteristik Der physiologische Prozess der Erzeugung menschlicher Sprache ist Teil der Standardliteratur zur Sprachsignalverarbeitung. An dieser Stelle soll bereits auf Tabelle 4.2 (Sonoritätsklassen) vorgegriffen werden. Sie beschreibt im Wesentlichen, dass stimmhafte Laute energiereicher als stimmlose sind. Die Stärke der Öffnung des Mundes bildet dabei eine Art Ordnung der abgestrahlten Lautstärke. Stimmhafte Laute haben ihre spektrale Ausprägung besonders im tiefen Frequenzbereich. Stimmlose sind breitbandig bzw. haben besondere Ausprägung im hohen Frequenzbereich. Daraus kann geschlussfolgert werden, dass der Mensch bei tiefen Frequenzen (besonders bei stimmhaften Lauten) energiereicheren Schall abstrahlt als bei hohen Frequenzen. Genauer kann hier darauf nicht eingegangen werden.

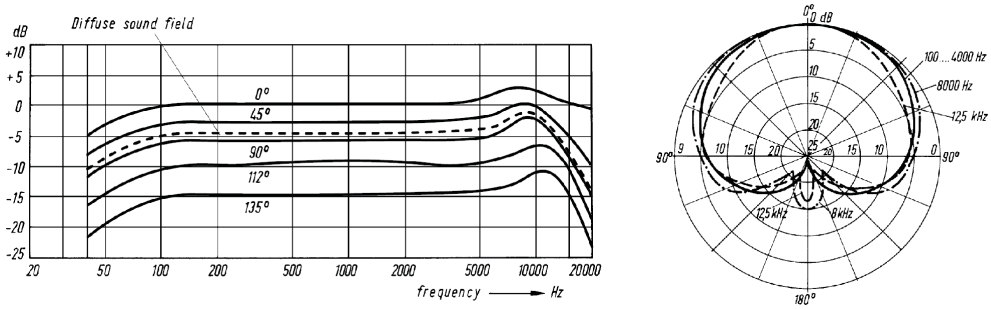

Abbildung 3.4 – Beispiel für einen Druckgradientenempfänger. Links: Frequenzgang. Rechts: Richtcharakteristik. Graphiken aus [BP99].

3.2.3.3 Mikrofon

Die vorliegende Arbeit widmet sich nicht dem Aufbau und der Wirkungsweise von Mikrofonen, dazu wird auf die Standardliteratur verwiesen. Dennoch sollen die Richtcharakteristik und das Übertragungsverhalten von Mikrofonen kurz beschrieben werden.

Richtcharakteristik Man unterscheidet bei Mikrofonen meist zwischen zwei Wirkprinzipien:

- Druckempfänger
- Druckgradientenempfänger

Druckempfänger haben eine Kugelcharakteristik und Druckgradientenempfänger eine 8-Charakteristik, werden aber meist mit einer Kugelcharakteristik überlagert, sodass die bekannte Nierenform (bzw. Abarten der Niere) entsteht (vgl. Abbildung 3.4 rechts). In Haushaltgeräten würde man aus Kostengründen Elektretkapseln als Mikrofone verbauen, wo beide Charakteristiken möglich sind [BP99].

Frequenzcharakteristik Mikrofone werden im Rahmen dieser Arbeit als lineare Systeme angenommen, die Schalldruck in eine Spannung wandeln. Dabei stellen sowohl die mechanischen Grenzen des Mikrofons als auch die Betriebsspannung natürliche Grenzen des Aussteuerbereichs dar. Gleiches gilt für den nachgeschalteten Mikrofonverstärker sowie den Analog-Digital-Umsetzer ADU. Auch hier ergeben sich Aussteuerungsgrenzen. In dieser Arbeit wird vorausgesetzt, dass der Aussteuerbereich nicht überschritten wird. Mikrofonhersteller sind bemüht, den Frequenzgang des Mikrofons konstant zu gestalten. Dies ist nicht immer vollständig möglich (vgl. Abbildung 3.4, links), soll aber für die weiteren Betrachtungen in dieser Arbeit angenommen werden. Zusätzlich wird noch auf den Nahbesprechungseffekt bei Druckgradientenempfängern

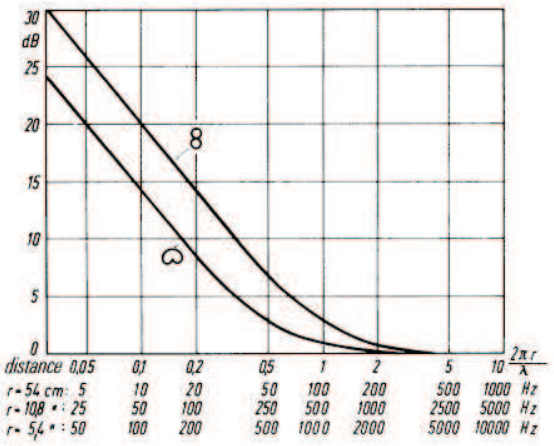

Abbildung 3.5 – Nahbesprechungseffekt bei einem Druckgradientenempfänger. Graphik aus [BP99].

verwiesen, der zu einer Anhebung von tiefen Frequenzen führt, wenn die Quelle sehr nahe am Mikrofon ist (Abbildung 3.5). Daher sind sie für Headsetmikrofone eher ungeeignet. Auch der Nahbesprechungseffekt wird im weiteren Verlauf ignoriert. Bei Kugelmikrofonen tritt er nicht auf und für Nierenmikrofone wird er vernachlässigt bzw. es wird pauschal angenommen, dass der Quelle-Mikrofon-Abstand so groß ist, dass er keine Wirkung hat.

3.3 Raumimpulsantworten

3.3.1 Sprecher-Raum-Mikrofon-System

Betrachtet man die Sprecher-Raum-Mikrofon-Strecke unter systemtheoretischen Gesichtspunkten, so ergibt sich ein System, das sämtliche akustischen Effekte in der mathematischen Formulierung seiner Übertragungsfunktion $G(j\omega)$ widerspiegelt. In der Akustik ist es üblich, vorzugsweise die Fourierrücktransformierte von $G(j\omega)$, die Impulsantwort $g(t)$, zu betrachten. Die Impulsantwort beschreibt ein System in allen (systemtheoretischen) Eigenschaften vollständig, wenn es sich dabei um ein lineares System handelt [WS06], was hier angenommen werden kann [Kut00].

Abbildung 3.6 zeigt ein vollständiges Modell des SRM-Systems. Die beiden Zweige auf der linken Seite sowie der Schalter symbolisieren die beiden unterschiedlichen Quellzweige Mensch (Index S (von Sprecher)) und Lautsprecher (Index LS). Da das Eingangssignal für die gedachte Anwendung gewöhnlich ein Sprachsignal ist, soll es

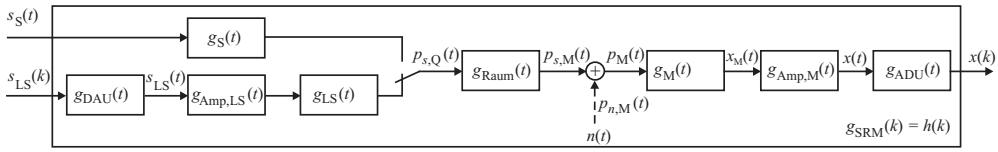

Abbildung 3.6 – Detailliertes Blockschaltbild des Sprecher-Raum-Mikrofon-Systems und des akustischen Szenarios.

die Notation $s(t)$ erhalten. Bei der menschlichen Anregung kann als Eingangssignal des Systems die gedachte Volumengeschwindigkeitsfunktion $Q(t)$ der Quelle bzw. deren Ableitung $\dot{Q}(t)$ gesehen werden, vgl. Gleichung (3.18). Es erhält hier die Notation $s_S(t)$. Es folgt ein System $g_S(t)$, welches die Produktion der menschlichen Sprache aus dem Quellsignal $s_S(t)$ symbolisch darstellen soll. $g_S(t)$ kann als einfacher Proportionalitätsfaktor beschrieben werden, der $s_S(t)$ in das Signal der Schalldruck-Zeit-Funktion von $s(t)$ an der Quelle $p_{s,Q}(t)$ umwandelt. Dieses wird dann vom eigentlichen Raum $g_{Raum}(t)$ zum Schalldrucksignal (herrührend von $s(t)$) am Mikrofon $p_{s,M}(t)$ gewandelt. Das Mikrofon stellt eine Additionsstelle zu anderen Signalen dar, die hier zusammenfassend als Geräusch $n(t)$ bezeichnet werden sollen. Das Geräusch steht am Mikrofon ebenfalls als Schalldruck-Zeit-Funktion $p_{n,M}(t)$ zur Verfügung. Das am Mikrofon eintreffende Gesamtsignal, die Schalldruck-Zeit-Funktion $p_M(t)$, wird vom Mikrofon $g_M(t)$ in das elektrische Signal $x_M(t)$ gewandelt. Der Mikrofonverstärker $g_{Amp,M}(t)$ wandelt $x_M(t)$ in das zeitkontinuierliche Eingangssignal zur Signalverarbeitung $x(t)$, welches vom Analog-Digital-Umsetzer (ADU) in das zeitdiskrete und quantisierte Eingangssignal für die Signalverarbeitung $x(k)$ überführt wird. Bei der menschlichen Anregung ist das Quellsignal $s_S(t)$ normalerweise nicht zugänglich. Es ist nur möglich, es nahe am Sprecher durch ein Mikrofon zu messen; man misst also das Signal $p_{s,Q}(t)$. Dies ist zwar dem Quellsignal $s_S(t)$ proportional, allerdings ist es in Reinform nicht messbar, da der Raum auch bei geringen Abständen von der Quelle bereits einen nicht unerheblichen verfälschenden Einfluss hat (vgl. Abbildung 3.2 bzw. 3.14, vgl. Erkennungsexperimente in Abschnitt 4). Wenn die Quelle durch einen Lautsprecher realisiert wird, ist das Eingangssignal das ebenfalls abgetastete und quantisierte Signal $s_{LS}(k)$. Es wird vom Digital-Analog-Umsetzer (DAU) in ein kontinuierliches Signal $s_{LS}(t)$ übertragen, das vom Verstärker und Lautsprecher in die Schalldruck-Zeit-Funktion an der Quelle $p_{s,Q}(t)$ gewandelt wird. Im Unterschied zur menschlichen Anregung ist hier das Eingangssignal zugänglich. Die Systeme $g_S(t)$, $g_{DAU}(t)$, $g_{Amp,LS}(t)$, $g_{LS}(t)$, $g_M(t)$, $g_{Amp,M}(t)$ sowie $g_{ADU}(t)$ werden zunächst als lineare Systeme angenommen, die rein proportionale Wirkung haben. D. h., sie verstärken, dämpfen oder überführen verschiedene physikalische Größen ineinander, ändern aber die Form des Signals selbst nicht. Aus diesen Annahmen heraus kann für all diese Systeme ein Gesamtproportionalitätsfaktor oder -übertragungsfaktor a_M definiert werden. Der Index M deutet darauf hin, dass dieser Proportionalitätsfaktor in die unterschiedliche Aussteuerung des Mi-

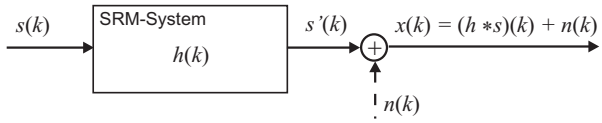

Abbildung 3.7 – Vereinfachtes Blockschaltbild des Sprecher-Raum-Mikrofon-Systems und des akustischen Szenarios.

krofonverstärkers modelliert werden kann, denn sie ist unbekannt und nicht fixiert; sie kann also von Mikrofon zu Mikrofon sehr unterschiedlich sein. Die Aussteuerbereiche von ADU, DAU, Verstärkern, Mikrofon und Lautsprecher sollen als eingehalten angenommen werden. Bei Übersteuerung eines dieser Systeme werden sie nichtlinear, wodurch zusätzliche Probleme enstehen, die in dieser Arbeit nicht diskutiert werden. Bei Einhaltung der Abtastbedingung kann ein kontinuierliches Signal, bzw. System, einfach zu den diskreten Zeitpunkten $t_k = k \cdot \Delta T_s$ beobachtet werden. Daraus entsteht die Zahlenfolge von k. Das Gesamtübertragungsverhalten kann also nunmehr einfach beschrieben werden durch

$$\begin{aligned} x(k) &= (g_{\mathrm{SRM}} * s)(k) + n(k) \\ &= a_{\mathrm{M}}(g_{\mathrm{Raum}} * s)(k) + n(k). \end{aligned} \tag{3.62}$$

Wie bereits angewandt, werden zeitkontinuierliche Systeme in der Regel durch die Funktionen $g(t)$ bzw. $G(j\omega)$ beschrieben. So definiert auch [Kut00] die Systeme Lautsprecher und Raum mit g bzw. G. Im Forschungsgebiet *blinde Enthallung* der Signalverarbeitung hat sich jedoch abweichend davon die Notation h bzw. H für das kontinuierliche SRM-System durchgesetzt. Darauf soll im Folgenden Bezug genommen werden, indem das SRM-System mit h bezeichnet wird. Aus (3.62) wird somit die bekannte Gleichung

$$x(k) = (h * s)(k) + n(k). \tag{3.63}$$

Nach diesen Annahmen kann das vereinfachte Blockschaltbild, wie es in Abbildung 3.7 dargestellt ist, definiert werden. Die gestrichelten Signalpfeile in Abbildung 3.6 und 3.7 deuten darauf hin, dass das Geräusch für diese Arbeit nicht betrachtet wird, es gilt

$$n(k) = 0. \tag{3.64}$$

Damit ergibt sich die recht einfache Beziehung

$$x(k) = (h * s)(k). \tag{3.65}$$

Bemerkung zu den Notationen: In der Signal- und Systemtheorie wird üblicherweise ein Übertragungsverhalten im Zeitbereich mit

$$y(t) = (g * x)(t) \tag{3.66}$$

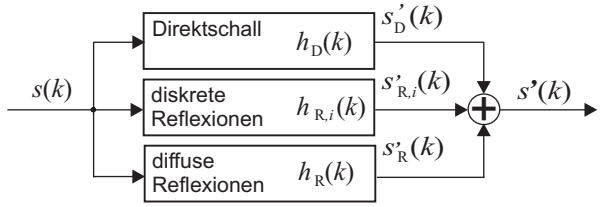

Abbildung 3.8 – Dekomposition des SRM-Systems in drei Subsysteme.

beschrieben. In dieser Arbeit wird jedoch $x(t)$ als Ausgangssignal des Systems $h(t)$ definiert. Die Gründe dafür liegen in der traditionellen Beschreibung von Spracherkennern, die $x(t)$ oder $x(k)$ gewöhnlicherweise als Eingangssignal definieren. In [VHH98] findet man ansonsten bspw. auch die Notation $s(k)$ für das Signal des Sprechers, $m(k)$ für das Mikrofonsignal oder $n(k)$ für das Geräusch, hier geht diese Arbeit teilweise konform.

Die Zeitfunktion $h(t)$ bzw. $h(k)$ soll von hier an auch als Raumimpulsantwort, kurz RIR (engl.: Room-Impulse-Response), bezeichnet werden.

3.3.2 Dekomposition in Subsysteme – Eigenschaften

Der Begriff Raumimpulsantwort ist etwas irreführend, da er impliziert, dass ein Raum genau eine Impulsantwort hat. Dem ist, wie den Ausführungen in den vorangegangenen Abschnitten entnehmbar, bekanntermaßen nicht so. Es existiert eine unendliche Anzahl von Raumimpulsantworten für einen Raum, die sich aus der unendlichen Anzahl von Quell- und Mikrofonpositionen ergibt. Diese Anzahl wird durch ein sinnvoll gewähltes Raster wieder endlich. Das Signal von Rasterpunkt zu Rasterpunkt unterscheidet sich jedoch erheblich. Die Anzahl an Rasterpunkten kann zudem noch mit allen möglichen Zuständen des Raumes multipliziert werden. Dabei spielen sowohl Gaszustände (wie Temperatur), geometrisch-räumliche Zustände (wie geöffnete Fenster, Möblierung etc.) oder Personen (Anzahl, Position, Kleidung etc.) eine Rolle. Es ist also zunächst nicht möglich, *die* Raumimpulsantwort zu bestimmen. Anstelle dessen scheint es für jeden Rasterpunkt eine völlig unterschiedliche Raumimpulsantwort zu geben. Aufgrund der Eigenschaften von direktem und diffusem Schallfeld gibt es dennoch Eigenschaften, die bei allen Positionen zumindest für einen Zustand identisch (idealisierte Annahme) oder zumindest ähnlich sind.

Direktes und diffuses Hallfeld spiegeln sich in der Raumimpulsantwort wider. Deshalb soll das SRM-System, wie in Abbildung 3.8, in drei Subsysteme unterteilt werden. Diese Unterteilung in drei Gruppen von Komponenten der RIR, Direktschall, diskrete Reflexionen und Nachhall (diffuse Reflexionen), ist eine klassische Betrachtung in der Raumakustik. Sie wird u. a. auch in den ursprünglichen Definitionen der raumakusti-

schen Maße (Abschnitt 4.4) verwendet, so z. B. in [SL74]. Das System der diskreten Reflexionen soll später, wie unten begründet, vernachlässigt werden, sodass ein System mit zwei Subsystemen verbleibt. Abbildung 3.9 zeigt beispielhaft eine gemessene Raumimpulsantwort in verschiedenen Ansichten. Die Eigenschaften der Subsysteme sind daran ebenfalls beobachtbar:

1. **Direkter Schall:** h_D beschreibt das direkte Schallfeld.
   - Es besteht aus einem einfachen Impuls. Daher entsteht ein breitbandiges (konstantes) Spektrum bei Kurzzeit-Spektralanalyse um den Impuls herum (vgl. Abbildung 3.9 (a)).
   - In der Impulsantwort erscheint der Impuls nach einer Totzeit t_0, die sich aus dem Abstand von Quelle und Mikrofon und aus der Schallgeschwindigkeit ergibt (in Abbildung 3.9 ist die Totzeit kaum erkennbar, da sie im Vergleich zum dargestellten Zeitfenster sehr gering ist.).
   - Die Amplitude ist proportional $\frac{1}{r}$ (vgl. Gleichung (3.27)).
   - Es gibt keine Frequenzabhängigkeit, da die Distanz wenige Meter beträgt (vgl. Tabelle 3.1 und Abschnitt 3.4.2).
2. **Diskrete Reflexionen:** $h_{\mathrm{R},i}$ beschreibt erste besonders starke Reflexionen, die an glatten Wänden etc. erzeugt werden. Sie treten in größeren Räumen deutlicher auf. In kleinen Räumen, wie im Wohnumfeld, fallen sie mit den diffusen Reflexionen nahezu zusammen (in Abbildung 3.9 (b) oder (c) sind ebenfalls keine diskreten Reflexionen zu beobachten). Deshalb wird dieses Subsystem im Folgenden vernachlässigt.
   - Es beschreibt einzelne Impulse am Anfang der Hallphase der RIR.
   - In der Impulsantwort erscheinen sie nach einer Totzeit, die sich aus der Gesamtstrecke $r_{\mathrm{R},i}$ (inklusive aller zurückgelegten Reflexionspfade) von der Quelle bis zum Mikrofon und der Schallgeschwingigkeit ergibt.
   - Ihre Amplitude berechnet sich aus der Abhängigkeit $\frac{1}{r_{\mathrm{R},i}}$ sowie dem Reflexionsgrad ρ der passierten Reflexionspunkte (vgl. Abschnitt 3.2.2).
   - Wegen der frequenzabhängigen Wandabsorption (vgl. Tabelle 3.2) sind diese Impulse im Gegensatz zu h_D nun frequenzabhängig.
3. **Diffuse Reflexionen - Hall:** Diffuse Reflexionen h_R entstehen durch die Wirkung des diffusen Hallfeldes. Sie beschreiben Schallwellen, die auf völlig unterschiedlichen Wegen und Reflexionspunkten zu verschiedenen Zeitpunkten am Mikrofon eintreffen.
   - Die Hallphase startet als weißes bis farbiges Rauschen (vgl. Abbildung 3.9 (a)).
   - Die mittlere Energie verringert sich nach einer Exponentialfunktion (vgl. Gleichung (3.39)). Das drückt sich auch in der Pegeldarstellung von Abbildung 3.9 (c) aus (vgl. Einhüllende), da die Exponentialfunktion im Logarithmus zu einer Geraden wird (Gleichung (3.40)).
   - Die Geschwindigkeit des Energieabfalls ist frequenzabhängig (vgl. Abbildung

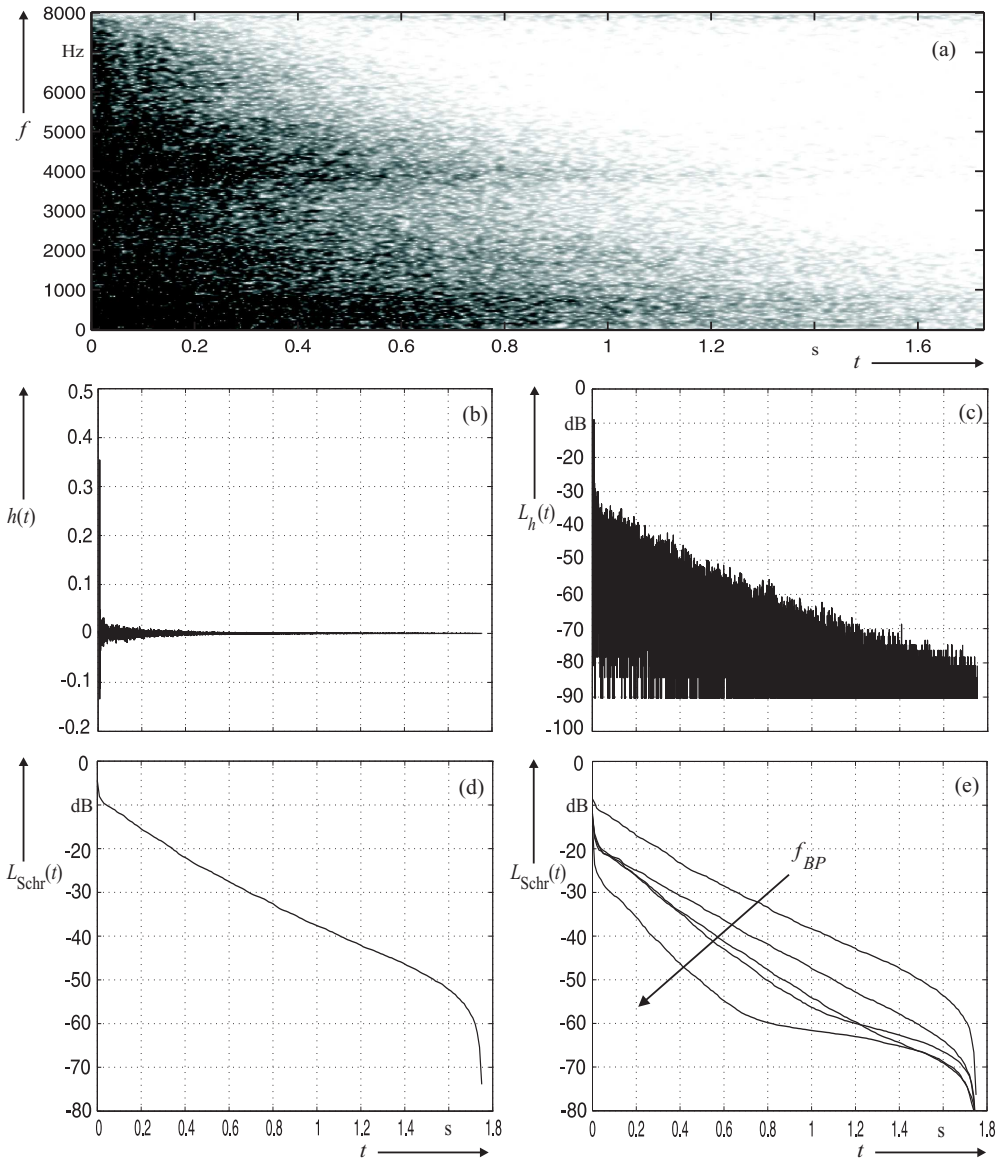

Abbildung 3.9 – Raumimpulsantwort gemessen in einem Treppenhaus ($T_{60} = 2$ s) in verschiedenen Darstellungsformen: (a) Spektrogramm, (b) Zeitsignal $h(t)$, (c) Leistungssignal in Pegeldarstellung $L_h = 10 \lg h^2(t)$, (d) Schröder-Integral und (e) Schröder-Integral in 5 Subbändern (f_{BP} - Mittenfrequenz des Bandpasses).

3.9 (a) und (e)), typisch ist ein Anstieg der Geschwindigkeit von tiefen zu hohen Frequenzen (vgl. auch die statistische Erhebung in Abschnitt 3.4.2). Daraus folgt zusätzlich eine frequenzabhängige Nachhallzeit. Es lassen sich zwei Gründe dafür angeben:
  - Wie in Abschnitt 3.2.1 dargestellt wurde, ist die Luftdämpfung bei größeren Distanzen nicht zu vernachlässigen. Man kann leicht ausrechnen, dass die Schallwellen des Nachhalls nach bspw. 300 ms bereits ca. 100 m zurückgelegt haben. Tabelle 3.1 gibt eine Vorstellung über die Stärke der Frequenzabhängigkeit bei diesen Distanzen.
  - Wie bereits bei $h_{\mathrm{R},i}$ erwähnt, ist der Reflexionsgrad ρ bzw. Absorptionsgrad α vom Material abhängig, wobei bei massiven Gebäuden tiefe Frequenzen tendenziell besser reflektiert werden als hohe (vgl. Tabelle 3.2).
- Die konkrete Funktion $h_{\mathrm{R}}(k)$ ist an jedem Raumpunkt unterschiedlich, aber der Energieabfall, bzw. die Funktion der mittleren Leistung, ist nicht von den Positionen von Mikrofon oder Quelle abhängig (Annahme in Gleichung (3.31)).

Daraus folgt, dass es, um akustische Abhängigkeiten eines bestimmten Raumes zu bestimmen, ausreichend ist, nur die Raumimpulsantworten zu messen, die eine Abhängigkeit vom Abstand zwischen Quelle und Mikrofon (SMD) beschreiben. Dabei ist es unerheblich, an welchen genauen Positionen im Raum sich Quelle und Mikrofon befinden, solange der dazwischenliegende SMD bekannt ist. Akustiker benutzen auch gern eine Raumimpulsantwort im Fernfeld, wo der Einfluss des direkten Schalles vernachlässigt werden kann. Hier wäre tatsächlich nur eine Impulsantwort nötig, um verschiedene Eigenschaften des Raumes zu ermitteln.

Mit der Dekomposition von $h(k)$ und der Überlegung, dass $h_{\mathrm{R},i}$ vernachlässigt werden kann, entsteht

$$h(k) = h_{\mathrm{D}}(k) + h_{\mathrm{R}}(k), \tag{3.67}$$

eine Beschreibung, wie sie bspw. auch in [Kut00, NKM07] vorkommt.

3.3.3 Modell einer Raumimpulsantwort

Raumimpulsantworten können mittels der in den vorangegangenen Abschnitten beschriebenen Eigenschaften modelliert werden. Hier wird das Modell der Subsysteme (Abbildung 3.8) herangezogen. Gleichung (3.67) dient als Ausgangspunkt. Um die Komponenten $h_{\mathrm{D}}(t)$ und $h_{\mathrm{R}}(t)$ zu berechnen, wird bereits jetzt auf Abschnitt 4.4.11 verwiesen, wo gezeigt wird, dass unter der Voraussetzung der Unkorreliertheit von $s(t)$ und $h(t)$ das Verhältnis der Leistungen von direktem und diffusem Signal ($\widetilde{x}^2_{\mathrm{D}}$ und $\widetilde{x}^2_{\mathrm{R}}$) aus dem Verhältnis der Energien von $h_{\mathrm{D}}(t)$ und $h_{\mathrm{R}}(t)$ ermittelt werden kann

$$\frac{\widetilde{x}^2_{\mathrm{D}}}{\widetilde{x}^2_{\mathrm{R}}} = \frac{W_{h_{\mathrm{D}}}}{W_{h_{\mathrm{R}}}}. \tag{3.68}$$

Weiterhin wird für das Verhältnis der Leistungen der Teilsignale das Verhältnis der Schallenergiedichten w_D bzw. w_R eingesetzt, das nach (3.58) mit

$$\frac{r_\mathrm{R}^2}{r^2} = \frac{w_\mathrm{D}}{w_\mathrm{R}} = \frac{\widetilde{x}_\mathrm{D}^2}{\widetilde{x}_\mathrm{R}^2} = \frac{W_{h_\mathrm{D}}}{W_{h_\mathrm{R}}} \tag{3.69}$$

ausgedrückt werden kann. Für den Signalaufbau soll noch das Gauß'sche, mittelwertfreie Rauschen $n_\mathrm{G}(t)$ definiert werden, das die Leistung

$$\tilde{n}_\mathrm{G}^2(t) = \sigma_\mathrm{G}^2 = 1 \tag{3.70}$$

besitzt.

Direktschallkomponente Die Direktschallkomponente der RIR kann (theoretisch) als ein Dirac-Impuls

$$h_\mathrm{D}(t) = b_\mathrm{D}\frac{1}{r}\delta(t) \tag{3.71}$$

beschrieben werden. Die SMD-abhängige Amplitude wird mit Gleichung (3.24) begründet. b_D ist ein Proportionalitätsfaktor, der die Stärke des Dirac-Impulses bestimmen soll. Er repräsentiert die Schallleistung der Quelle (3.23). Dies ist für eine genaue Simulation des Szenarios wichtig. Wird dabei ein sauberes Sprachsignal (im reflexionsfreien Raum) mit einem bestimmten SMD r aufgenommen, dann ist b_D so zu bestimmen, dass $W_{h_\mathrm{D}}(r) = 1$ wird. Somit wird der originale Schalldruck einbezogen. Ist der genaue Schalldruckpegel des verhallten Signals nicht von Bedeutung (Normalfall), da er letztlich von einem Verstärker etc. geregelt wird, so kann b_D beliebig gewählt werden, z. B. $b_\mathrm{D} = 1$. Die Energie W_{h_D} ergibt sich zu

$$\begin{aligned} W_{h_\mathrm{D}} &= \int_{-\infty}^{\infty} h_\mathrm{D}^2(t)\mathrm{d}t && (3.72)\\ &= \int_{-\infty}^{\infty} b_\mathrm{D}^2\frac{1}{r^2}\delta^2(t)\mathrm{d}t && (3.73)\\ &= b_\mathrm{D}^2\frac{1}{r^2} && (3.74)\\ &\sim \frac{1}{r^2}. && (3.75) \end{aligned}$$

Anstelle des Dirac-Impulses kann auch das oben definierte Gaußrauschen $b_\mathrm{G}n_\mathrm{G}(t)$ mit der Dauer t_D und der Leistung $\tilde{n}_\mathrm{G}^2 = b_\mathrm{G}^2$ verwendet werden. Die Energie dieses Rauschens muss dem des Dirac-Imulses aus (3.75) entsprechen

$$\begin{aligned} b_\mathrm{D}^2\frac{1}{r^2} &= \int_0^{t_\mathrm{D}} b_\mathrm{G}^2 n_\mathrm{G}^2\mathrm{d}t && (3.76)\\ &= b_\mathrm{G}^2 t_\mathrm{D}. && (3.77) \end{aligned}$$

Damit kann b_{G}^2 bestimmt werden und es entsteht

$$h_{\mathrm{D}}(t) = \begin{cases} b_{\mathrm{D}} \frac{1}{r} \frac{1}{\sqrt{t_{\mathrm{D}}}} n_{\mathrm{G}}(t) & ; \quad 0 \leq t \leq t_{\mathrm{D}} \\ 0 & ; \quad t < 0 \;\; ; \;\; t > t_{\mathrm{D}}. \end{cases} \tag{3.78}$$

t_{D} muss hinreichend kurz sein, um das Signal nicht zu verfälschen.

Diffusschallkomponente Die Beendigung des Direktschallimpulses stellt ein Abschalten der Schallquelle dar. Der entstehende Hall bildet ein Rauschen

$$n_h(t) = b_{\mathrm{R}} \cdot n_{\mathrm{G}}(t) \tag{3.79}$$

mit einem exponentiellen Energieabfall nach Gleichung (3.39). Damit ergibt sich in (3.39) die Anfangsenergie zu

$$w_{h_{\mathrm{R}}}(t_0) = \mathrm{E}\left\{n_h^2(t_0)\right\} = \mathrm{E}\left\{b_{\mathrm{R}}^2 n_{\mathrm{G}}^2(t_0)\right\} = b_{\mathrm{R}}^2. \tag{3.80}$$

Die Diffusschallkomponente der RIR

$$h_{\mathrm{R}}(t) = \begin{cases} b_{\mathrm{R}} n_{\mathrm{G}}(t) \cdot e^{-\frac{6,9}{T_{60}} t} & ; \quad t \geq 0 \\ 0 & ; \quad t < 0 \end{cases} \tag{3.81}$$

erfüllt Gleichung (3.39) nach Erwartungswertbildung ihrer Leistungsfunktion

$$w_{h_{\mathrm{R}}}(t) = \begin{cases} \mathrm{E}\left\{h_{\mathrm{R}}^2(t)\right\} = \mathrm{E}\left\{b_{\mathrm{R}}^2 n_{\mathrm{G}}^2(t) \cdot e^{-\frac{13,8}{T_{60}} t}\right\} = b_{\mathrm{R}}^2 \cdot e^{-\frac{13,8}{T_{60}} t} & ; \quad t \geq 0 \\ 0 & ; \quad t < 0. \end{cases} \tag{3.82}$$

Die Energie der Diffusschallkomponente ergibt sich demnach zu

$$W_{h_{\mathrm{R}}} = \int_{-\infty}^{\infty} h_{\mathrm{R}}^2(t) \mathrm{d}t \tag{3.83}$$

$$= \int_0^{\infty} b_{\mathrm{R}}^2 n_{\mathrm{G}}^2(t) \cdot e^{-\frac{13,8}{T_{60}} t} \mathrm{d}t = \left[-b_{\mathrm{R}}^2 \frac{T_{60}}{13,8} \cdot e^{-\frac{13,8}{T_{60}} t} \right]_0^{\infty} \tag{3.84}$$

$$= b_{\mathrm{R}}^2 \frac{T_{60}}{13,8}. \tag{3.85}$$

Es ist noch der energiebezogene Faktor b_{R} zu ermitteln. Er bestimmt, in welchem Verhältnis Direktschallenergie und Diffusschallenergie zusammenhängen. Dazu werden die Gleichungen (3.85) und (3.74) in (3.69) eingesetzt

$$\frac{r_{\mathrm{R}}^2}{r^2} = \frac{b_{\mathrm{D}}^2 \frac{1}{r^2}}{b_{\mathrm{R}}^2 \frac{T_{60}}{13,8}} \tag{3.86}$$

und man erhält

$$b_{\mathrm{R}} = \frac{b_{\mathrm{D}}}{r_{\mathrm{R}}}\sqrt{\frac{13,8}{T_{60}}}. \tag{3.87}$$

Mit $b_{\mathrm{D}} = 1$ wird damit (3.81) zu

$$h_{\mathrm{R}}(t) = \begin{cases} \frac{1}{r_{\mathrm{R}}}\sqrt{\frac{13,8}{T_{60}}} n_{\mathrm{G}}(t) \cdot e^{-\frac{6,9}{T_{60}}t} & ; \quad t \geq 0 \\ 0 & ; \quad t < 0 \end{cases} \tag{3.88}$$

und (3.82) zu

$$w_{h_{\mathrm{R}}}(t) = \begin{cases} \frac{1}{r_{\mathrm{R}}^2}\frac{13,8}{T_{60}} \cdot e^{-\frac{13,8}{T_{60}}t} & ; \quad t \geq 0 \\ 0 & ; \quad t < 0 \end{cases} \tag{3.89}$$

und (3.85) zu

$$W_{h_{\mathrm{R}}} \quad = \quad \frac{1}{r_{\mathrm{R}}^2}. \tag{3.90}$$

Vollständiges Modell Nach Einsetzen von $b_{\mathrm{D}} = 1$ in (3.71) (was, wie erwähnt, für Simulationszwecke möglich ist) entsteht das Modell der RIR, welches sich aus

$$h_{\mathrm{D}}(t) \quad = \quad \frac{1}{r}\delta(t) \tag{3.91}$$

$$h_{\mathrm{R}}(t) \quad = \quad \begin{cases} \frac{1}{r_{\mathrm{R}}}\sqrt{\frac{13,8}{T_{60}}} \cdot n_{\mathrm{G}}(t) \cdot e^{-\frac{6,9}{T_{60}}t} & ; \quad t \geq 0 \\ 0 & ; \quad t < 0 \end{cases} \tag{3.92}$$

zusammensetzt[6]. Alternativ besteht die Möglichkeit, h_{D} nach (3.78) darzustellen. Eine Schwäche des hier beschriebenen Modells ist die fehlende Einbeziehung der Frequenzabhängigkeit von T_{60}. Allerdings ist das Modell auf eine Subbanddarstellung

[6] Gleichung (3.81) wird oft als die bekannte stochastische Approximation einer Raumimpulsantwort nach Schröder bezeichnet. In der Forschungsgemeinde wird üblicherweise [Sch81] als Quelle angegeben. Bei genauer Recherche findet man diese Form jedoch bereits ein Jahr zuvor in Houtgast et al. [HSP80]. Aus diesem Grund bemerkt Schröder in [Sch81], dass sein Paper bereits zwei Jahre vorher von der Veröffentlichung zurückgewiesen wurde; nur ein Abstract [Sch78] wurde gedruckt, der allerdings auch von [HSP80] referenziert wird. Dennoch fällt auf, dass die Definition (3.81) in [Sch81] nur am Rande behandelt wird. Sowohl der energiebezogene Faktor b_{R} als auch die Direktschallkomponente sind nicht berücksichtigt. Schröder macht zwar keinen Fehler, da er (3.81) als Approximation der Hallphase bezeichnet, welche für das Fernfeld auch als vollständige RIR angenommen werden kann, allerdings sind die Ausführungen in [HSP80] wesentlich genauer, da sie auch das Direktschallfeld mit einbeziehen. Im Unterschied zu [Sch81] definiert [HSP80] anstelle der Signalverläufe die Energieverläufe von (3.91) und (3.92). Eine Herleitung der Gleichungen findet man auch in [HSP80] nicht. Dennoch bezieht sich diese Arbeit eher auf die Ausführungen in Houtgast und Steeneken [HSP80].

erweiterbar, was diesen Nachteil aufheben würde. Des Weiteren ist anzumerken, dass es schwierig ist, eine realistische Abhängigkeit von nur einem Parameter zu erhalten. Ändert sich bspw. die Nachhallzeit, so ändert sich bei konstant gehaltenem Volumen auch der Hallradius, bei konstant gehaltenem Hallradius das Volumen etc. r_R ist durch Richtcharakteristik, Volumen und Nachhallzeit bestimmt (Gleichungen (3.51), (3.52)). Unter der Annahme, dass die Richtcharakteristik vernachlässigt wird, besteht also immer noch der Parameter Volumen, der ebenfalls festzulegen ist, um eine SMD-Abhängigkeit zu beschreiben. D. h., es ist nur möglich, einen angenommenen Raum mit bekannter Größe und Nachhallzeit auf diese Weise abzubilden. Die modellierte Impulsantwort ist dann nur für ein bestimmtes Volumen gültig. Man kann allerdings mit (3.47) zeigen, dass dieselbe Impulsantwort auch für einen Raum mit gleicher T_{60}, aber anderem Volumen anwendbar ist. Sie gilt dennoch bei geändertem SMD r_2

$$r_2^2 = r_1^2 \frac{V_2}{V_1}. \tag{3.93}$$

Derartig generierte Impulsantworten wurden u. a. für Experimente zum Verhalten von Sprachverständlichkeit [HSP80], von Spracherkennung [SGK06, LUA08] oder von F_0-Detektionsmethoden [UH08c] unter Hallbedingungen benutzt. Allerdings wird dabei meist nur die Fernfeldapproximation (3.88) in die Simulation einbezogen (abgesehen von [SGK06] arbeiten alle anderen Genannten nur mit der Fernfeldapproximation).

3.3.4 Künstliches Verändern der Nachhallzeit

Hat man zu Simulationszwecken nur eine bestimmte Anzahl von gemessenen Raumimpulsantworten zur Verfügung, die eine gesuchte Nachhallzeit gerade nicht enthalten, so ist es möglich, daraus eine passende Raumimpulsantwort zu generieren. Dazu wird die Hallphase mit einer Exponentialfunktion bewertet

$$h_\mathrm{R,neu}(t) \quad = \quad h_\mathrm{R,alt}(t) \cdot e^{-\frac{1}{\tau_D}}. \tag{3.94}$$

Die Dämpfungskonstante τ_D kann durch Hinzuziehen des Modells in (3.81) bestimmt werden

$$b_\mathrm{R} n_\mathrm{G}(t) \cdot e^{-\frac{6,9}{T_\mathrm{60,neu}} t} = b_\mathrm{R} n_\mathrm{G}(t) \cdot e^{-\left(\frac{6,9}{T_\mathrm{60,alt}} + \frac{1}{\tau_D}\right) t} \tag{3.95}$$

$$\frac{6,9}{T_\mathrm{60,neu}} = \frac{6,9}{T_\mathrm{60,alt}} + \frac{1}{\tau_D} \tag{3.96}$$

$$\tau_D = \frac{T_\mathrm{60,alt} T_\mathrm{60,neu}}{T_\mathrm{60,alt} - T_\mathrm{60,neu}}. \tag{3.97}$$

Mit dieser Verfahrensweise kann $T_\mathrm{60,neu}$ sowohl kleiner als auch größer als $T_\mathrm{60,alt}$ gewählt werden. Eine zu starke Vergrößerung ist jedoch praktisch nicht möglich, da

die gemessene Raumimpulsantwort $h_{\mathrm{R,alt}}(t)$ nicht unendlich exponentiell abfällt, sondern zum Zeitpunkt t_N in ein Rauschen mündet (reale Messung). Nimmt man an, dass die Veränderung von T_{60} dazu dient, den gleichen Raum mit vergrößerter äquivalenter Absorptionsfläche, wie bei geöffnetem Fenster oder einer größeren Anzahl an Polstermöbeln, zu simulieren, so bleibt zwar das Volumen gleich, aber der Hallradius vergrößert sich (Gleichungen (3.51) bzw. (3.52)).

3.3.5 Bestimmung der Nachhallzeit – Schröder-Integral

Wie erwähnt, ist die Nachhallzeit die wichtigste Größe in der Raumakustik. Sie kann prinzipiell immer dann ermittelt werden, wenn eine Schallquelle verstummt, da dann nach (3.11) und (3.40) die Geschwindigkeit des Abfalls des Schalldruckpegels gemessen werden kann. Ein Vergleich verschiedener Methoden zur Ermittelung der Nachhallzeit wird in [VB94] vorgestellt. Zur Messung der Nachhallzeit empfiehlt [ISO97] zwei Methoden:

- Interrupted-Noise-Method und
- Integrated-Impulse-Response-Method

Bei der ersten Methode wird der Raum mit einem Rauschen angeregt und nach Abschalten wird anhand des Abfalls der Energie die Nachhallzeit gemessen. Im Rahmen dieser Arbeit wurde vorwiegend die zweite Methode verwendet. Sie folgt dem weit verbreiteten Verfahren der Rückwärtsintegration der Raumimpulsantwort, welches 1965 von M. R. Schröder vorgestellt wurde [Sch65]. An der Impulsantwort kann die Nachhallzeit beobachtet werden, wenn ihre Pegeldarstellung betrachtet wird. Abbildung 3.9 (c) zeigt, dass die Einhüllende der Hallphase eine fallende Gerade ist, an deren Anstieg die Nachhallzeit abgelesen werden kann. Jedoch kann dies aufgrund des unruhigen Verlaufs zu Ungenauigkeiten führen. Schröder hatte die Idee, die verbleibende Energie der Raumimpulsantwort zum Zeitpunkt τ durch Integration der Leistungskurve $x^2(t)$ nach

$$L_{\mathrm{Schr}}(\tau) = 10\lg\left(\int_{\tau}^{\infty} x^2(t)\mathrm{d}t\right)\mathrm{dB} \tag{3.98}$$

zu ermitteln. Nach Logarithmierung entsteht der sehr ruhige Verlauf von $L_{\mathrm{Schr}}(\tau)$, wie er in Abbildung 3.9 (d) dargestellt ist. Praktisch benutzt man die diskrete Variante der Summation, wobei die Summe nicht bis ins Unendliche, sondern bis zum Endindex K der Impulsantwort gebildet wird. Die durch die vielen Summationen entstehende lange Rechenzeit kann verkürzt werden, indem man die Schrittweite Δ_κ zwischen den Summationen auf mehrere Samples (z. B. $\Delta_\kappa = 100$) erhöht

$$L_{\mathrm{Schr}}(\kappa) = 10\lg\left(\sum_{k=\Delta_\kappa\cdot\kappa}^{K} x^2(k)\right)\mathrm{dB}. \tag{3.99}$$

Die Nachhallzeit ergibt sich, wenn an einer idealisierten Geraden das Zeitintervall κ_{60} abgelesen wird, bei dem $L_{\mathrm{Schr}}(\kappa)$ um 60 dB gefallen ist. Sie ergibt sich nach

$$T_{60} = \frac{\Delta_\kappa \cdot \kappa_{60}}{f_s}. \tag{3.100}$$

Da in der Praxis Störgeräusche auftreten, hat man den SNR von 60 dB oft nicht zur Verfügung. So empfiehlt es sich, den Abfall nur für die ersten 20 oder 30 dB zu betrachten und die Gerade dann zu extrapolieren.

3.4 Messung von Raumimpulsantworten

3.4.1 Messmethoden

Um die Raumimpulsantwort bei einem bestimmten SMD zu ermitteln, ist das System mit einem Impuls anzuregen. Dies wurde früher auch in Form von Pistolenschüssen, Starterklappen etc. so gehandhabt. Nahezu ideale Impulse sind technisch allerdings schwierig zu realisieren. Das gilt besonders, wenn man sie mit einem Lautsprecher erzeugen möchte. Besser ist es, mit einem Lautsprecher ein definiertes Anregungssignal abzuspielen. Dabei gibt es die folgenden beiden Möglichkeiten zur Bestimmung der Impulsantwort, wovon die erste mit logarithmischem Gleitsinus als Anregungssignal für die Messungen in dieser Arbeit verwendet wurde.

3.4.1.1 Bestimmung aus der Übertragungsfunktion aus Eingangs- und Ausgangssignal

Wird der Raum von einem Lautsprecher mit einem Frequenzgemisch angeregt, in dem alle relevanten Frequenzen enthalten sind, kann die Impulsantwort über die Übertragungsfunktion aus Eingangs- und Ausgangssignal nach

$$\underline{H}(n) = \frac{\mathrm{DFT}\{x(k)\}}{\mathrm{DFT}\{s_{\mathrm{LS}}(k)\}}, \tag{3.101}$$

$$h(k) = \mathrm{IDFT}\{\underline{H}(n)\} \tag{3.102}$$

berechnet werden. Durch den Hall wird $x(k)$ länger als das Eingangssignal. Da die beiden DFTs die gleiche Dimension haben müssen, wird $s_{\mathrm{LS}}(k)$ mit Nullen aufgefüllt (Zero-Padding). Des Weiteren sollten Anfang und Ende der beiden Signale auf 0 enden, um Sprungstellen an den Signalgrenzen zu vermeiden. Dafür kann ein einfaches Ein- und Ausblendfenster benutzt werden.

Als Anregungssignal sind verschiedene Formen möglich. Es müssen alle relevanten Frequenzen enthalten sein; sie müssen aber nicht gleichmäßig verteilt sein[7]:

- **Linearer Gleitsinus** (nach [Hof98]) Ein Gleitsinus (auch Chirp oder Sweep) ist eine Sinusfunktion, deren momentane Frequenz $f_M(\tau)$ sich mit der Zeit langsam ändert

$$x(t) = \sin\left(2\pi \int_0^t f_M(\tau) \cdot \mathrm{d}\tau\right). \tag{3.103}$$

Die Zeitvariable τ wurde gewählt, um sie als Integrationsvariable von der oberen Integrationsgrenze t zu unterscheiden. Beim linearen Sweep folgt die Änderung einer linearen Kennlinie

$$f_M(\tau) = \frac{f_{g,o} - f_{g,u}}{T_{\mathrm{Sweep}}}\tau + f_{g,u}. \tag{3.104}$$

$f_{g,o}$ und $f_{g,u}$ bezeichnen die obere und untere Grenzfrequenz und T_{Sweep} die Sweepdauer. Das Argument im Sinus berechnet sich damit

$$2\pi \int_0^t f_M(\tau) \cdot \mathrm{d}\tau = 2\pi \left[\frac{f_{g,o} - f_{g,u}}{2T_{\mathrm{Sweep}}}\tau^2 + f_{g,u}\tau\right]_0^t \tag{3.105}$$

und der Sweep ergibt sich demnach zu

$$x(t) = \sin\left(2\pi\left(\frac{f_{g,o} - f_{g,u}}{2T_{\mathrm{Sweep}}}t^2 + f_{g,u}t\right)\right). \tag{3.106}$$

- **Logarithmischer Gleitsinus** (nach [Hof98]) Beim logarithmischen Sweep ändert sich die Momentanfrequenz nach einer Exponentialfunktion

$$f_M(\tau) = a \cdot e^{b\cdot\tau}. \tag{3.107}$$

Nach Lösen des Gleichungssystems

$$f_{g,u} = a \cdot e^{b\cdot 0} \tag{3.108}$$

$$f_{g,o} = a \cdot e^{b\cdot T_{\mathrm{Sweep}}} \tag{3.109}$$

erhält man für die Momentanfrequenz die Funktion

$$f_M(\tau) = f_{g,u} \cdot e^{\frac{T_{\mathrm{Sweep}}}{\ln\left(\frac{f_{g,o}}{f_{g,u}}\right)}\cdot\tau}. \tag{3.110}$$

Nach Einsetzen in Gleichung (3.103) und Lösung des Integrals ergibt sich der Sweep zu

$$x(t) = \sin\left(2\pi \cdot f_{g,u} \cdot \frac{T_{\mathrm{Sweep}}}{\ln\left(\frac{f_{g,o}}{f_{g,u}}\right)} \cdot \left(e^{\frac{T_{\mathrm{Sweep}}}{\ln\left(\frac{f_{g,o}}{f_{g,u}}\right)}\cdot t} - 1\right)\right). \tag{3.111}$$

[7] Eine interessante Gegenüberstellung verschiedener Anregungssignale findet man in [STFP96].

- **Multisinus** Ein Gemisch aus mehreren Sinussignalen, die den relevanten Frequenzbereich abdecken sollen, wird als Multisinus bezeichnet.
- **Weißes Rauschen** Weißes Rauschen enthält gleichverteilt alle Frequenzen und ist damit ebenfalls gut geeignet.

3.4.1.2 Korrelationsmethode

Schreibt man die diskrete Faltung in Gleichung (3.65) aus, entsteht

$$x(k) = (h * s)(k) = \sum_{l=-\infty}^{\infty} h(l) \cdot s(k-l). \tag{3.112}$$

Die Kreuzkorrelation von zwei Signalen $x(k)$ und $s(k)$ ist definiert nach

$$\psi_{sx}(m) = \lim_{K\to\infty} \frac{1}{2K+1} \sum_{k=-K}^{K} s(k) \cdot x(k+m). \tag{3.113}$$

Für den Sonderfall, dass $s(k) = x(k)$ ist, entsteht $\psi_{ss}(m)$, die Autokorrelationsfunktion. Wenn man $x(k+m)$ in (3.113) durch (3.112) substituiert, folgt

$$\psi_{sx}(m) = \lim_{K\to\infty} \frac{1}{2K+1} \sum_{k=-K}^{K} s(k) \cdot \sum_{l=-\infty}^{\infty} h(l) \cdot s(k+m-l) \tag{3.114}$$

und durch Umordnen

$$\psi_{sx}(m) = \sum_{l=\infty}^{\infty} h(l) \left[\lim_{K\to\infty} \frac{1}{2K+1} \sum_{k=-K}^{K} s(k) \cdot s(k+m-l) \right]. \tag{3.115}$$

Der Teil in den eckigen Klammern ergibt die Autokorrelationsfunktion $\psi_{ss}(m-l)$, sodass man schließlich

$$\psi_{sx}(m) = \sum_{l=-\infty}^{\infty} h(l)\psi_{ss}(m-l) \tag{3.116}$$

schreiben kann, oder auch kurz

$$\psi_{sx}(m) = (h * \psi_{ss})(m). \tag{3.117}$$

Diese Beziehung ist aus der Systemtheorie bekannt. Wenn man ein Eingangssignal $s(k)$ findet, dessen Autokorrelationsfunktion ein Dirac-Impuls ist

$$\delta(k) = \psi_{ss}(k), \tag{3.118}$$

dann ist die Kreuzkorrelationsfunktion des Ausgangs- und Eingangssignales gerade die Impulsantwort des Systems

$$h(k) = \psi_{sx}(m) = (h * \delta)(m). \tag{3.119}$$

Ein solches Eingangssignal ist bspw. ein Rauschen oder idealerweise eine sogenannte Maximalfolge (zur Erzeugung von Maximalfolgen vgl. z. B. [VHH98]). Benutzt man ein solches Signal als Anregungssignal $s(k)$ in (3.112), kann mit dem aufgezeichneten Ausgangssignal $x(k)$ die Kreuzkorrelationsfunktion nach (3.113) berechnet werden und es ergibt sich nach (3.119) die RIR $h(k)$.

3.4.1.3 Kompensation des Lautsprechers

In Abschnitt 3.3.1 wird zunächst angenommen, dass alle Teilsysteme aus Abbildung 3.6 mit Ausnahme des Teilsystems $g_{\text{Raum}}(t)$ (das eigentliche akustische System Raum) als simple Proportionalitätsfaktoren modelliert werden können. [Kut00] empfiehlt allerdings noch die Kompensation des Lautsprechers. Dazu wird der Frequenzgang des ermittelten Systems noch mit dem Kehrwert des Frequenzganges des Lautsprechers $\underline{G}_{\text{LS}}^{-1}(n)$ multipliziert.

$$\underline{H}_{\text{komp}}(n) = \underline{H}_{\text{mess}}(n) \cdot \underline{G}_{\text{LS}}^{-1}(n) \tag{3.120}$$

Der Frequenzgang des Lautsprechers kann mit dem gleichen Verfahren bestimmt werden, mit dem auch das SRM-System gemessen wird. Allerdings ist dazu ein reflexionsfreier Raum erforderlich.

3.4.2 Statistische Analyse der Wohn- und Büroumgebung

Um herauszufinden, welche Wertebereiche der Verhallung im Anwendungsszenario Haushaltsgeräte mit Sprachsteuerung, der Wohn- und Büroumgebung, zu erwarten sind, wurde im Rahmen der Arbeit eine Studie durchgeführt, in der in mehreren Räumen Raumimpulsantworten systematisch aufgezeichnet und ausgewertet wurden. Gemessen wurde mit einem kompakten und mobilen Messaufbau, bestehend aus einem Aktivlautsprecher mit 10 cm Tieftöner und Hochtonkalotte, einem PC-Elektret-Tischmikrofon und einem Notebook mit Soundkarte für Signalein- und -ausgabe. Die Studie umfasste Wohn- und Bürogebäude, wobei insgesamt die Raumklassen (Wohnzimmer, Küche, Bad und Büro) mit jeweils 20 Beispielen vertreten waren. Innerhalb eines Raumes wurden Raumimpulsantworten, soweit möglich, für die festen SMDs $r = (50; 100; 200; 300; 400; 600; 800)$ cm aufgezeichnet. Ist der Raum kleiner, wird als letztes abweichend von dem vorgegebenen Raster beim größtmöglichen SMD gemessen. Aufgrund der Größen der Wohn- und Büroräume waren somit mindestens drei Impulsantworten pro Raum messber, meist aber vier bis fünf. Gesteuert wurden die

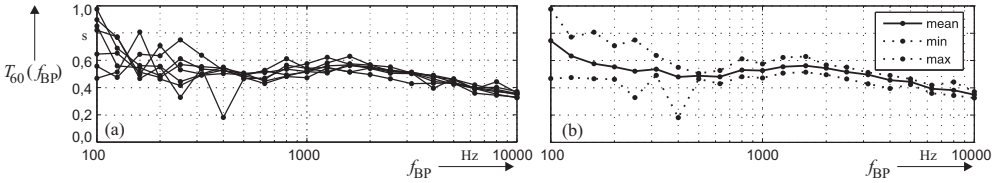

Abbildung 3.10 – Nachhallzeit eines Beispielraumes gemessen in Subbändern. Es handelt sich in diesem Beispiel um ein großes Wohnzimmer, deshalb konnten auch für alle vorgegebenen SMDs Raumimpulsantworten gemessen werden. (a) Messung der Nachhallzeit in Subbändern $T_{60}(f_{\mathrm{BP}})$ bei den einzelnen SMDs $r = (50;\ \ldots;\ 800)$ cm. (b) Statistische Auswertung der Einzelmessung von (a). Es werden der arithmetische Mittelwert (mean) von $T_{60}(f_{\mathrm{BP}})$ bei jedem f_{BP} gebildet sowie die obere (max) und untere (min) Einhüllende von (a) dargestellt.

Versuche mit der Messsoftware *DIRAC* von der Firma Bruel und Kjaer [BK08], mit der auch die Nachhallzeit in Subbändern $T_{60}(f_{\mathrm{BP}})$ gemessen werden konnte. Die Mittenfrequenzen der Subbänder lagen bei $f_{\mathrm{BP}} = (100; 125; 160; 200; 250; 315; 400; 500; 630; 800; 1000; 1250; 1600; 2000; 2500; 3150; 4000; 5000; 6300; 8000)$ Hz. Die Messergebnisse eines Beispielraumes beschreibt Abbildung 3.10. Zunächst ist zu beobachten, dass die Nachhallzeit für einen einzelnen Raum (auch frequenzbezogen $T_{60}(f_{\mathrm{BP}})$) positionsunabhängig ist. Besonders bei den sehr tiefen Frequenzen bis 250 Hz sind größere Schwankungen der Messergebnisse zu beobachten, das liegt u. a. daran, dass bei diesen Wellenlängen Eigenmoden des Raumes auftreten (stehende Wellen, deren Energie von Ort zu Ort verschieden ist). Deshalb wird die Mittelung der Einzelmessungen durchgeführt, vgl. Abbildung 3.10 (b).

Aus diesen Messungen der Einzelräume konnte eine Statistik über alle Räume gebildet werden. Sie sollte Aufschluss darüber geben, ob sich die vier Raumklassen in ihren akustischen Eigenschaften unterscheiden. Abbildung 3.11 zeigt die Ergebnisse für die vier Raumklassen. Es wurde zunächst über die arithmetischen Mittelwerte der Einzelräume (mean aus Abbildung 3.10 (b)) für alle Räume einer Klasse gemittelt (mean of means). Da die Einzelräume sich teilweise erheblich unterscheiden, wurden auch der Maximal- (max of means) und Minimalwert (min of means) der arithmetischen Mittelwerte sowie noch die jeweiligen Extremwerte aus allen einzelnen Messungen (max of maxs/ min of mins) aufgezeichnet, um damit ein vollständiges Bild über die Verteilung der Nachallzeiten in der jeweiligen Raumklasse zu erhalten. Der Vollständigkeit halber ist noch die obere und untere Standardabweichung (upper/lower stddev) in die Diagramme eingetragen.

Die Unterschiede zwischen den Raumklassen sind nicht signifikant. Die Ergebnisse stellen jedoch deutlich dar, dass die Nachhallzeit, wie vorher theoretisch beschrieben, frequenzabhängig ist und in der Tendenz für diese Räume von tiefen zu hohen Fre-

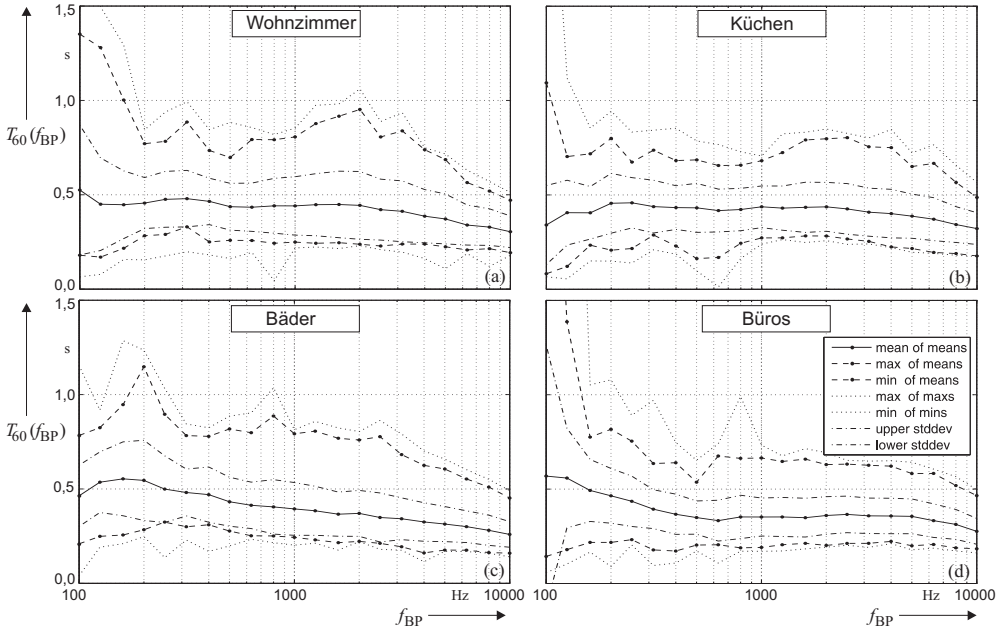

Abbildung 3.11 – Statistische Auswertung von T_{60} gemessen in Subbändern mit der Mittenfrequenz des filternden Bandpasses f_{BP} bei den verschiedenen Raumklassen.
• **mean of means** Mittelwert aller Mittelwerte der Einzelräume.
• **max of means** bzw. **min of means** Maximal- bzw. Minimalwert aller Mittelwerte der Einzelräume (beschreibt gemessenen Bereich der Stichproben).
• **max of maxs** bzw. **min of mins** Maximal- bzw. Minimalwert aller Einzelmessungen (Extremlimits).
• **upper stddev** bzw. **lower stddev** (mean of means) + bzw. − Standardabweichung
.

quenzen abnimmt[8].

Grenzwerte für die Nachhallzeit Die Gesamtnachhallzeit in den einzelnen Räumen ist in Abbildung 3.12 (a) dargestellt. Tabelle 3.3 wertet die Stichproben statistisch mit arithmetischem Mittelwert μ und Standardabweichung σ aus. Wenn man eine Normalverteilung voraussetzt, deuten die Werte $\mu + \sigma$ (84,15 % aller Werte liegen darunter)

[8] Diese Tendenz ist bei modernen Bauten ohne massives Mauerwerk (z. B. große Glasflächen) nicht unbedingt gegeben. Aber die Statistik zeigt hier auch, dass Wohn- und Büroräume tendenziell eher Massivbauten sind. Einschränkung: Messungen wurden in Deutschland durchgeführt, in anderen Ländern bzw. Kulturkreisen variiert die Bauweise; in Japan ist bspw. auch eine sehr leichte Holzbauweise stark verbreitet. Zudem ist die Anzahl der Gebäude von 20 noch sehr gering, um diese Tendenz zu verifizieren.

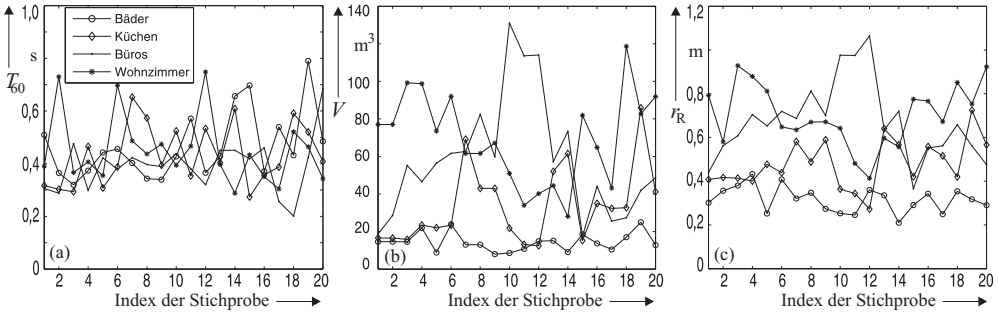

Abbildung 3.12 – Verteilung der ermittelten Größen bei den vier Raumklassen Bad, Küche, Büro und Wohnzimmer. (a) Nachhallzeit. (b) Volumen. (c) Hallradius (nach (3.52) berechnet. Annahme $\gamma = 1$). Der Raumindex auf der Abszisse stellt die Nummer der Stichprobe in der jeweiligen Raumklasse dar.

sowie $\mu + 2\sigma$ (97,7 % aller Werte liegen darunter) die oberen Grenzen der Nachhallzeiten an. Es können keine raumklassenspezifischen Unterschiede in den Nachhallzeiten festgestellt werden. Aber es kann deutlich gezeigt werden, dass Räume im Wohn- und Bürobereich Nachhallzeiten zwischen 0,3 s und 0,8 s besitzen. Diese Aussage ist sehr wichtig für die Experimente und Ansätze, die in den folgenden Kapiteln behandelt werden.

Grenzwerte für den SMD Eine weitere wichtige Aussage aus diesen Messungen betrifft den SMD. Die Räume wurden zusätzlich in Länge, Breite und Höhe vermessen. Tabelle 3.3 wertet die Ergebnisse statistisch aus. Man kann den maximalen SMD anhand der statistisch ermittelten Längen bestimmen bzw. über die Berechnung der Diagonalen aus Längen und Breiten. Aussagekräftiger für einen maximalen SMD ist allerdings eher die Breite als Länge oder gar Diagonale. Man stelle sich ein Wohnzimmer vor, in dem an der Längsseite ein sprachbedienter Fernseher steht. Der maximale SMD ergibt sich aus dem Abstand gegenüberliegender Wände minus einem Meter (Grobschätzung), da im Normalfall weder Gerät (zumindest nicht das Mikrofon im Gerät) noch Sprecher direkt an der Wand stehen. Dieser Abstand bildet den Radius eines gedachten Halbkreises, innerhalb dessen sich ein Sprachbediener intuitiv bewegen würde[9]. Die Anzahl von 20 gewählten Stichproben ist auch hier noch zu klein, um die Aussagen sicher zu verallgemeinern; auch die Normalverteilung kann nicht zwingend angenommen werden. Dennoch stellen die Ergebnisse in Tabelle 3.3 anschaulich die Größenordnungen dar, in denen sich der SMD bewegt. Unter Beachtung der Breite findet man den maximalen SMD etwa bei (4 . . . 5) m im untersuchten Anwendungsszenario (Das Bad bildet hier eine Ausnahme, was auch im folgenden Abschnitt dargestellt

[9]Diese Aussage ist heuristisch. Sie basiert auf Erfahrungen.

Tabelle 3.3 – Statistische Auswertung zu den Nachhallzeiten und Abmaßen (Länge, Breite, Volumen; die Höhe spielt für den SMD keine Rolle, sie würde auch innerhalb eines Gebäudes nicht und von Gebäude zu Gebäude nur wenig variieren). μ – arithmetischer Mittelwert, σ – Standardabweichung. Die Limits $[\mu - \sigma,\ \mu + \sigma]$ und $[\mu - 2\sigma,\ \mu + 2\sigma]$ deuten die Bereiche an, in denen 68.3 % bzw. 95.4 % aller Stichproben erwartet werden, wenn eine Gaußverteilung angenommen wird. Demnach werden $(68.3 + \frac{31.7}{2})\% = 84.15\%$ bzw. $(95.4 + \frac{4.6}{2})\% = 97.7\%$ aller Stichproben unterhalb der betreffenden oberen Limits erwartet. Die Eignung einer Gaußverteilung für diesen Zufallsprozess wird hier nicht weiter überprüft, da es sich nur um eine Grobabschätzung von Bereichen handelt. Für die berechneten Obergrenzen wird daher eine Unsicherheit in Kauf genommen; dennoch ergeben sie bereits gute Richtwerte, mit denen zunächst gearbeitet werden kann.

Raumklasse	Bad	Küche	Büro	Wohnzimmer	gesamt
T_{60} μ	0,46 s	0,43 s	0,39 s	0,45 s	0,43 s
T_{60} σ	0,12 s	0,11 s	0,10 s	0,13 s	0,12 s
T_{60} $\mu + \sigma$	0,59 s	0,55 s	0,49 s	0,58 s	0,55 s
T_{60} $\mu + 2\sigma$	0,72 s	0,66 s	0,59 s	0,71 s	0,67 s
Länge μ	2,94 m	4,05 m	5,08 m	6,53 m	4,66 m
Länge σ	0,55 m	1,30 m	1,32 m	1,76 m	1,83 m
Länge $\mu + \sigma$	3,49 m	5,35 m	6,39 m	8,30 m	6,48 m
Länge $\mu + 2\sigma$	4,04 m	6,65 m	7,71 m	10,06 m	8,31 m
Breite μ	1,99 m	3,05 m	3,63 m	4,08 m	3,21 m
Breite σ	0,45 m	0,94 m	1,24 m	0,85 m	1,18 m
Breite $\mu + \sigma$	2,44 m	3,99 m	4,87 m	4,93 m	4,39 m
Breite $\mu + 2\sigma$	2,89 m	4,93 m	6,11 m	5,77 m	5,56 m
Volumen μ	14,41 m^3	33,81 m^3	56,32 m^3	69,43 m^3	43,99 m^3
Volumen σ	4,87 m^3	20,44 m^3	34,34 m^3	24,37 m^3	30,90 m^3
Volumen $\mu + \sigma$	19,28 m^3	54,25 m^3	90,66 m^3	93,80 m^3	74,90 m^3
Volumen $\mu + 2\sigma$	24,16 m^3	74,70 m^3	125,00 m^3	118,17 m^3	105,80 m^3

wird). Auch diese Aussage ist sehr wichtig für die folgenden Kapitel.

Volumen und Hallradius Laut Gleichung (3.36) ist die Nachhallzeit vom Volumen abhängig. Mit einer vereinfachend angenommenen Quaderform konnte somit mit den gemessenen Abmaßen auch eine Statistik der Volumina erhoben werden, die in Abbildung 3.12 (b) dargestellt ist. Sie zeigt, dass Räume im Wohn- und Geschäftsbereich ebenfalls auch typische Raumgrößen zwischen 20 und 100 m^3 besitzen. Erwartungsgemäß kann bei den Volumina jedoch eine Klassenspezifik festgestellt werden. Auch hier gibt Tabelle 3.3 eine statistische Auswertung. Wohnzimmer und Büros sind in der Tendenz größer als Küchen; Bäder sind durchgängig am kleinsten. Das entspricht der normalen Wahrnehmung der Umwelt und wird hier zusätzlich einmal statistisch

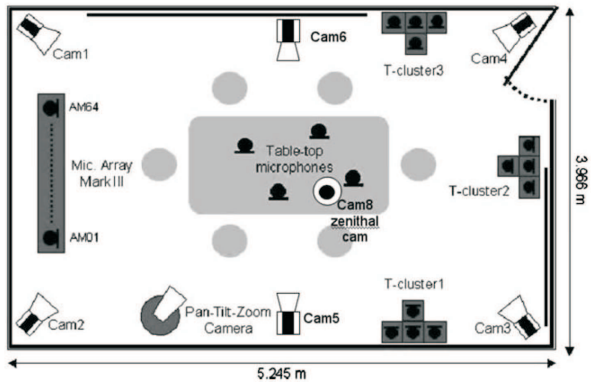

Abbildung 3.13 – Skizze des SMART-Rooms. Dargestellt sind drei T-Mikrofon-Arrays, ein 64-Mikrofon-Line-Array, fünf Kameras, ein Tisch mit Stühlen und vier Tischmikrofonen. Graphik entnommen aus [NGMH07].

belegt. Ebenfalls aus der täglichen Wahrnehmung würde man erwarten, dass Bäder besonders hallig sind, was aufgrund ihrer Nachhallzeit in Abbildung 3.12 (a) bzw. Tabelle 3.3 jedoch nicht nachgewiesen werden kann. Dennoch besitzen Bäder typischerweise eine geflieste Oberfläche, parallele Wände und wenig absorbierendes Material wie Teppiche, Polstermöbel o. ä. Die äquivalente Absorptionsfläche sollte also demnach gering sein, was mit Gleichung (3.36), den gemessenen Nachhallzeiten und Volumina dargestellt werden kann (Zur Verteilung der äquivalenten Absorptionsfläche wird hier keine Abbildung dargestellt.). Aus diesen Werten kann auch der Hallradius nach (3.51) berechnet werden, vgl. Abbildung 3.12 (c)[10]. Es ist nun ersichtlich, dass die Bäder durchgängig einen geringeren Hallradius besitzen, wodurch sich die tägliche Wahrnehmung der besonderen Halligkeit bestätigt; in normaler Kommunikationsentfernung von etwa (1 ... 2) m ist man bereits weit im diffusen Hallfeld (vgl. auch Abbildung 3.2, wo der Mittelwert der gemessenen Bäder eingezeichnet wurde). Aus den frequenzabhängigen Nachhallzeiten und den Volumina lässt sich ebenfalls eine sehr interessante Frequenzabhängigkeit der Hallradien berechnen (Die zugehörige Graphik ist hier nicht dargestellt.).

3.4.3 Messungen in der SMART-Room-Umgebung

Im Rahmen des von der Europäischen Union geförderten Projektes CHIL (Computers In the Human Interaction Loop [CHIL]) wurde u. a. an der UPC (Universitat Politècnica de Catalunya, Barcelona, Spanien) ein sogenannter SMART-Room ein-

[10] Die Richtcharakteristik nach (3.52) ist dabei vernachlässigt worden. Demnach können die Graphen noch mit einem Proportionalitätsfaktor zwischen 1 ($\gamma = 1$) und 4 ($\gamma = 16$) gestreckt werden.

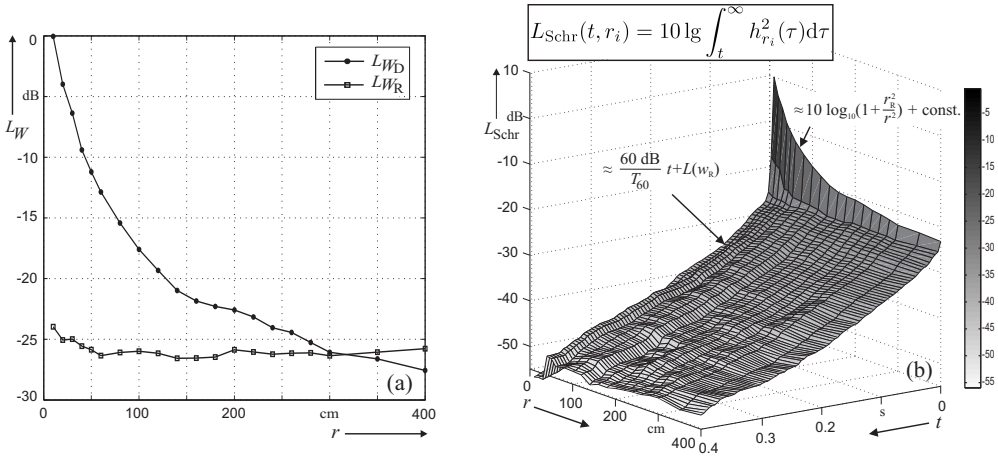

Abbildung 3.14 – Auswertung der im SMART-Room [NGMH07] an der UPC in Barcelona gemessenen RIRs mit steigendem SMD r. (a) Pegeldarstellung der Energien W_D und W_R (vgl. Gleichungen (4.52) und (4.53), der (nicht ideale) Impuls wird mit $t_k = 5$ ms von der Hallphase unterschieden). Diese Graphik ist laut Abschnitt 4.4.11 auch zur Darstellung der Energie des Direktschalls w_D im Vergleich zur Energie des Hallfeldes w_R geeignet. Bei etwa 300 cm ist der Hallradius als Schnittpunkt der beiden Graphen angedeutet. (b) Schröder-Integral unter der zusätzlichen Abhängigkeit vom SMD. In beiden Graphiken ist die Positionsunabhängigkeit der diffusen Schallenergie w_R gut zu beobachten. Für beide Graphiken gelten die gemessenen SMDs $r = (10; 20; 30; 40; 50; 60; 80; 100; 120; 140; 160; 180; 200; 220; 240; 260; 280; 300; 350; 400)$ cm.

gerichtet. Er ist quaderförmig und hat die Größe: (3966, 5245, 2800) mm (Breite, Länge, Höhe). Die Wände bestehen aus glatten Spanplatten. Der Raum ist spärlich eingerichtet, demnach sehr hallig. Er simuliert einen Beratungsraum (vgl. Abbildung 3.13). Das Szenario dient zu Forschungszwecken auf den Gebieten der Audiosignalverarbeitung: Automatic Speech Recognition, Speaker Identification, Speech Activity Detection, Acoustic Source Localization, Acoustic Event Detection und Speech Synthesis. Zusätzlich wird die Kombination mit Videotechniken untersucht. Im Rahmen der vorliegenden Arbeit wurden verschiedene Messungen im SMART-Room durchgeführt, u. a. eine positionsabhängige Messung der Raumimpulsantwort. Die Nachhallzeit wurde mit ca. (650 ± 50) ms ermittelt (Diese etwas ungenaue Angabe rührt daher, dass das Schröder-Integral keine exakte Gerade bildet, sondern eine leichte Krümmung hat; im Folgenden wird mit $T_{60} = 0,7$ s gearbeitet.). Laut Gleichung (3.51) berechnet sich mit dem Volumen von 58,2 m^3 und einer Nachhallzeit von 0,7 s der Hallradius bei Quelle und Mikrofon mit Kugelcharakteristik zu 0,51 m. Im Gegensatz zu den Messungen in Abschnitt 3.4.2 wurde versucht, die Richtcharakteristik und die Mundhöhe des Menschen grob mit einzubeziehen, indem der Lautsprecher während der Messungen vor

der Brust gehalten wurde. In Abbildung 3.14 (a) kann ein Hallradius von etwa 3 m festgestellt werden. Mit (3.52) kann daraus ein Richtfaktor $\gamma = 34$ berechnet werden. Diese Richtcharakteristik ist etwas zu stark. Eine Erklärung dafür ist die Überlagerung der Richtcharakteristiken von Lautsprecher und Mensch. Die Abschattung des menschlichen Körpers vor der Brust ist zudem größer als am Kopf. Ein Kunstkopf, der derzeit die gängige Simulation der menschlichen Abstrahlung darstellt, stand jedoch für die Messungen nicht zur Verfügung.

In Abbildung 3.14 (b) wird das Schröder-Integral über die SMD-abhängigen Raumimpulsantworten gebildet. Diese Graphik beinhaltet zwei sehr interessante Aussagen. Erstens ist ersichtlich, dass der Direktschall positionsabhängig ist. Da im Schröder-Integral zum Zeitpunkt $t = 0$ die Gesamtenergie der RIR erscheint, entsteht eine Proportionalität nach (3.56). Bei Pegelbildung ergibt sich die Funktion

$$L_{\mathrm{Schr}}(0, r) = 10 \lg \left(1 + \frac{r_{\mathrm{R}}^2}{r^2}\right) + \text{const.} \tag{3.121}$$

Die zweite Aussage besagt, dass der Hall annähernd positionsunabhängig ist. Dies kann durch den SMD-unabhängigen Anstieg der fallenden Gerade beobachtet werden.

3.5 Zusammenfassung der Erkenntnisse

In diesem Kapitel wurde die raumakustische Umgebung von Spracherkennern für die Wohn- und Büroumgebung beschrieben. Dabei wurden zunächst relevante physikalische Aspekte vorgestellt. Die Raumimpulsantwort wird als raumbeschreibende Funktion eingeführt. Die meisten raumakustischen Phänomene können an ihr beobachtet werden. Eine Studie zur Messung von Raumimpulsantworten über eine Anzahl von Räumen gibt Aufschluss über die zu erwartenden Bedingungen für den Einsatz im Wohn- und Büroumfeld. Folgende Erkenntnisse konnten gewonnen werden:

- Die Stärke der Hallstörung kann anhand eines DRR beschrieben werden.
- Der DRR verbessert sich, wenn die Schallquelle bzw. das Mikrofon eine Richtcharakteristik besitzt.
- Der DRR ist abhängig von T_{60}, V und SMD. Für einen festen Raum ist der DRR nur vom SMD, also von der Position des Sprechers abhängig.
- Der Hallradius beschreibt den Übergang zwischen Nah- und Fernfeld. Er vereint die beiden Raumgrößen T_{60} und V. Im Nahfeld und etwas über den Hallradius hinaus spielt das direkte Schallfeld eine wichtige Rolle. Der Hallradius ist in diesem Bereich eine wichtige raumakustische Größe. Für ideale Kugelcharakteristiken von Quelle und Mikrofon hat er im Wohn- und Büroumfeld etwa die Größenordnung (30 ... 100) cm. Bei Quellen und Senken mit Richtcharakteristik erhöhen sich diese Grenzen um bis zu Faktor 4 ($\gamma = 16$).

- Im Fernfeld spielt für die Akustik nur noch T_{60} eine Rolle. Im Wohn- und Büroumfeld liegt sie in den Grenzen $0,3\ \mathrm{s} \leq T_{60} \leq 0,8\ \mathrm{s}$.
- Der maximale SMD im Wohn- und Büroumfeld beträgt etwa $(4 \ldots 5)$ m.
- T_{60} ist frequenzabhängig. Gebäude in Massivbauweise tendieren dazu, dass T_{60} zu hohen Frequenzen hin geringer wird.

4 Die Wirkung des Raumes auf Sprache und Spracherkennung

4.1 Überblick

Nachdem die raumakustische Umgebung ausgiebig für die hier relevanten Zwecke beschrieben wurde, soll dieses Kapitel den Einfluss des Raumes auf die Spracherkennung herausfinden. Dies beginnt zunächst mit der Beschreibung von Sprachsignalen selbst, wobei sich die Darstellung auf die für die vorliegende Arbeit wichtigen Aspekte beschränkt. Anschließend wird dargestellt, welchen Einfluss Hall auf Sprachsignale ausübt. Beginnend mit der Definition traditioneller Maße wird versucht, ein Maß für die Störung des Sprachsignals durch Hall zu finden. Es folgen Experimente zur Spracherkennung unter verschiedenen raumakustischen Umgebungsbedingungen, die den stark störenden Einfluss von Raumhall auf die Spracherkennung darstellen. Durch gezielte Manipulation der Hallphasen von Raumimpulsantworten kann durch weitere Experimente festgestellt werden, welche Bereiche des Halls besonders störend wirken und welche größtenteils harmlos sind.

4.2 Ausgewählte Aspekte menschlicher Sprachsignale

Um den Einfluss des Raumes auf die Sprache zu untersuchen, werden zunächst hier relevante Begriffe und Definitionen eingeführt, die Sprachsignale beschreiben.

Die Einzelheiten der physiologischen Prozesse der menschlichen Sprachsignalerzeugung findet man in der medizinischen Literatur, aber auch üblicherweise in Standardwerken der Sprachsignalverarbeitung, z. B. in [VHH98, BSH08a]. In diesem Abschnitt wird deshalb der Fokus auf die für diese Arbeit wichtigen Eigenschaften gelegt.

4.2.1 Spracherzeugung, Grundfrequenz, Formanten

Anregungsarten Die Erzeugung des menschlichen Sprachsignals wird üblicherweise durch das Quelle-Filter-Modell beschrieben [Fan60]. Dabei wird zwischen den zwei Anregungsarten unterschieden:

- **stimmhaft** - Ausatmen von Luft aus den Lungen durch die Stimmritze, wobei die Stimmlippen die Stimmritze periodisch öffnen und schließen. Es entsteht ein periodisch unterbrochener Luftstrom mit der Grundperiode T_0 und deren Kehrwert, der Grundfrequenz F_0, der den Schall des stimmhaften Anregungssignals $e_v(t)$ (engl.: voiced Excitation) beschreibt. $e_v(t)$ ist ein periodisches, nicht sinusförmiges Signal, das als eine Folge verschliffener Impulse beschrieben werden kann. Demnach beinhaltet es Oberschwingungen (Harmonische) bei Vielfachen von F_0 und lässt sich demnach für einen angenommenen stationären Bereich in eine Fourierreihe entwickeln. Die Amplituden der Harmonischen sind die entsprechenden Fourierkoeffizienten.
- **stimmlos** - Ausatmen von Luft aus den Lungen bei weit geöffneter Stimmritze, wobei durch Verengen (bei Frikativen) oder plötzlichem Öffnen (bei Plosiven) an lautspezifischen Stellen im Mund-Rachen-Raum Luftverwirbelungen erzeugt werden. Es entsteht eine rauschartige Schallanregung der Luft $e_u(t)$ (engl.: unvoiced Excitation).

Zusätzlich gibt es Laute mit Mischanregung, etwa bei einem stimmhaften Frikativ (z. B. /z/ in *sauber*) oder Plosiv (z. B. /g/ in *ganz*). Das Sprachsignal $s_{\mathrm{Ph}}(t)$ des entsprechenden Lautes (der Index Ph bezieht sich auf Phon (Laut)) ergibt sich dann durch

$$s_{\mathrm{Ph}}(t) = e_{\mathrm{Ph}}(t) * v_{\mathrm{Ph}}(t), \tag{4.1}$$

wobei die Impulsantwort $v_{\mathrm{Ph}}(t)$ den Vokaltrakt (Mund-Rachen-Raum) beschreibt. Der entsprechende Laut wird im Wesentlichen durch die Übertragungsfunktion $\underline{V}_{\mathrm{Ph}}(t)$ beschrieben. Bspw. kann ein stimmhafter Laut mit unterschiedlichen bzw. variierenden, die Tonhöhe bestimmenden Grundfrequenzen gesprochen werden. Aus der Faltung wird im Frequenzbereich die Multiplikation

$$\underline{S}_{\mathrm{Ph}}(\omega) = \underline{E}_{\mathrm{Ph}}(\omega) \cdot \underline{V}_{\mathrm{Ph}}(\omega). \tag{4.2}$$

Man kann ein Sprachsignal auch als Überlagerung von harmonischen und nichtharmonischen Komponenten, $s_h(t)$ bzw. $s_n(t)$, zu einem bestimmten Zeitpunkt t betrachten

$$s(t) = s_h(t) + s_n(t). \tag{4.3}$$

Beim harmonischen Anteil handelt es sich um eine Addition von Sinusfunktionen bei Vielfachen der Grundfrequenz, herrührend von stimmhafter Anregung. Beim nichtharmonischen Anteil handelt es sich um Rauschen, herrührend von Verwirbelungen an Engstellen des Vokaltrakts bei stimmhafter oder stimmloser Anregung. Die Form und Gewichtung der beiden Komponenten hängt vom gerade gesprochenen Laut ab.

Grundfrequenz F_0 Nach [Hes08] bewegt sich die Grundfrequenz beim Menschen in einem Bereich von

$$50 \text{ Hz} \leq F_0 \leq 800 \text{ Hz}. \tag{4.4}$$

Tabelle 4.1 – Abhängigkeiten der statistischen Verteilung der Grundfrequenz vom Geschlecht. Die Studie von Ohara [Oha99] zeigt, dass die Grundfrequenz zusätzlich noch spezifisch zur gesprochenen Sprache bzw. vorhandenen Kultur ist.

	männlich	weiblich
Wertebereich nach Honda [Hon08]		
$F_{0,\min}$	80 Hz	120 Hz
$F_{0,\max}$	400 Hz	800 Hz
Mittelwerte nach Ohara [Oha99]		
für englische Sprecher $\overline{F}_{0,\text{eng.}}$	120 Hz	220 Hz
für japanische Sprecher $\overline{F}_{0,\text{jpn.}}$	150 Hz	340 Hz
Mittelwerte aus Datensätzen in Abschnitt 7.3		
für englische Sprecher (ein Sprecher pro Geschlecht) $\overline{F}_{0,\text{eng.}}$	100 Hz	205 Hz
für japanische Sprecher (14 Sprecher pro Geschlecht) $\overline{F}_{0,\text{jpn.}}$	130 Hz	235 Hz

Für die untere Grenze findet man bei [Hon08] einen unterschiedlichen Wert ($F_{0,\min} = 80$ Hz, vgl. Tabelle 4.1). Die angegebenen Grenzen stellen bereits Extremwerte dar, die bei normaler Sprache selten erreicht werden. Für die vorliegende Arbeit werden daher Limits in den Algorithmen auf

$$\begin{array}{rcrcrcr} F_{0,\min} & = & 70 \text{ Hz} & \text{bzw.} & T_{0,\max} & = & 14,28 \text{ ms} \\ F_{0,\max} & = & 600 \text{ Hz} & \text{bzw.} & T_{0,\min} & = & 1,67 \text{ ms} \end{array} \tag{4.5}$$

gesetzt, mit der Grundperiode T_0, die sich aus dem Reziproken von F_0 bestimmt. Eine Änderung der Grundfrequenz kann nicht sprungartig, sondern nur gleitend erfolgen. Die Grundfrequenz ist vom Geschlecht und vom Alter abhängig. Männer haben tiefe Stimmen. Frauen haben hohe Stimmen. Kinder können noch höhere Stimmen haben. Zusätzlich gibt es eine Abhängigkeit von kulturellen Gegebenheiten [Oha99], vgl. Tabelle 4.1. Vgl. auch Abbildung 7.4 zur Verteilung der Grundfrequenz.

Formanten Im Kurzzeitspektrum der meisten Laute bilden sich Maxima bei Resonanzfrequenzen des Vokaltrakts. Diese Maxima werden als Formanten F_1, F_2, F_3, F_4 usw. bezeichnet. Die Formanten liegen für die deutsche Sprache etwa in den Frequenzbereichen $F_1 \approx (200 \ldots 700)$ Hz, $F_2 \approx (800 \ldots 2000)$ Hz (wenige Ausnahmen gehen bis 2300 Hz) und $F_3 \approx (2500 \ldots 2900)$ Hz (vgl. Abbildung 4.1). F_1, F_2 und F_3 bestimmen den gesprochenen Vokal; F_4 (und höhere Formanten) charakterisiert eher den Sprecher und ist für das Verständnis des Lautes nicht von Bedeutung [Pom95, Lad95]. Diese Angaben gelten für statische Phasen von Lauten (isolierte Langvokale).

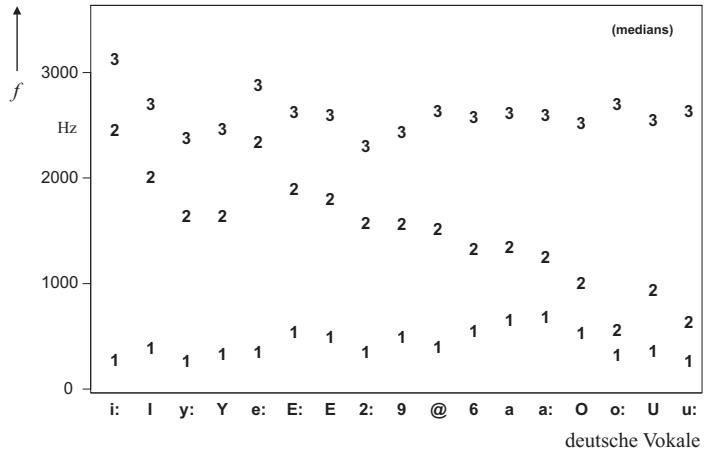

Abbildung 4.1 – Formantfrequenzen von Vokalen im Deutschen (statische Vokale, Langvokale). Die Graphik ist eine von B. Möbius erstellte Umsetzung einer tabellarischen Darstellung in [Moe01] und erscheint in seinen Vorlesungsunterlagen an der Universität Stuttgart.

4.2.2 Lautstärkepegel, Lautheit, Sonoritätsklassen

In Abschnitt 3.2.1 werden zur Beschreibung der Energie von Schallereignissen das Effektivwertquadrat des physikalisch messbaren Schalldrucks $\tilde{p}^2$ bzw. der zugehörige Schalldruckpegel L_p eingeführt. Die Lautstärke ist ein psychoakustisch motivierter Begriff, der beschreibt, wie stark ein Schallereignis vom Menschen wahrgenommen wird. Die beschreibende Größe ist der Lautstärkepegel, gemessen in Phon. Der Lautstärkepegel entspricht bei der Frequenz von 1000 Hz exakt dem Schalldruckpegel. Die Wahrnehmung des Menschen ist frequenzabhängig. Für Frequenzen $\neq 1000$ Hz beschreiben die sogenannten Kurven gleicher Lautstärkepegel (Isophone) die Zugehörigkeit zum Schalldruckpegel bei der entsprechenden Frequenz [DIN72, Kut04].

Der Lautstärkepegel hat den Nachteil, dass z. B. eine Verdopplung der Lautstärkeempfindung nicht durch eine Verdopplung des zugehörigen Pegels beschrieben wird. Deshalb wird die Größe Lautheit, gemessen in Sone, eingeführt [ZF90]. Die Lautheit folgt etwa einer Exponentialfunktion des Lautstärkepegels, d. h., eine lineare Steigerung des Lautstärkepegels hat einen exponentiellen Anstieg der Lautheit zur Folge.

Die Sonorität ist ein Begriff aus der Phonologie. Sie beschreibt die Schallfülle eines Lautes, die mit der Signalenergie in Bezug gebracht werden kann. Die einzelnen Laute besitzen unterschiedliche Sonorität, wobei der Sprecher durch seine Sprechweise Einfluss auf die Betonung und Lautheit eines Lautes hat. D. h., es ist nicht ohne Weiteres möglich festzulegen, dass ein /a/ mehr Energie besitzt als ein /r/. In der

Tabelle 4.2 – Sonoritätsklassen von Lauten im Deutschen. Die Werte wurden aus [Mei98] übernommen.

Sonoritätsoberklasse	Beispiele (SAMPA-Notation)	Sonorität
1. Vokale	a	10
	e, o	9
	i, u	8
2. Liquidkonsonanten	r	7
	l	6
3. Nasalkonsonanten	m, n, N	5
4. Frikativkonsonanten	s	4
	v, z, S	3
	f, h, C	2
5. Plosivkonsonanten	b, d, g	1
	p, t, k	0,5

Tendenz, also im Mittel über eine Menge an Stichproben, wird dies dennoch so sein. Ein Schema, das Regeln für die anzunehmenden Lautstärken von Lauten herstellt, ist das Prinzip von Sonoritätsklassen. Dabei wird zunächst in Sonoritätsoberklassen unterschieden, deren Klassenelemente einen gleichen Sonoritätswert haben [Mei98]. Eine feinere Unterteilung wird von Selkrik vorgeschlagen [Sel84] (zitiert in [Mei98]). Sie geht von einer elfstufigen Unterteilung aus (Tabelle 4.2). Die Sonorität der einzelnen Laute hängt vom Vorhandensein von Stimme (stimmhaft/stimmlos) und zusätzlich vom Grad der Behinderung des Luftstromes ab [Mei98]. Als Beispiel kann der Unterschied von /a/ und /i/ in Tabelle 4.2 in Augenschein genommen werden. Nach dem Grad der Mundöffnung werden /a/-Laute als offene Vokale und /i/-Laute als geschlossene Vokale bezeichnet [Pom95]. Bei Plosivlauten und Frikativen gibt es z. B. einen Unterschied von Stimmhaftigkeit und Stimmlosigkeit, wie bei /f/ und /v/ (wie in *Fass* und *was*) oder /g/ und /k/ (wie in *gleich* und *Kleid*).

4.2.3 Zeitliche Struktur

Zeitliche Einhüllende – TMEs In Abschnitt 4.2.1 wird bereits auf das Quelle-Filter-Modell verwiesen, allerdings wird das Filter nur für quasistationäre Sprachsegmente, z. B. innerhalb eines Lautes, als weitgehend konstant angenommen. Betrachtet man nun Sprache über einen Laut hinaus, so ändert sich über eine Äußerung hinweg sowohl das Anregungssignal als auch das Filter. Die Änderung des Filters geschieht durch die Bewegung der sogenannten Artikulatoren, also des Mundes, der Zunge usw. Durch die zeitliche Änderung des Filters werden verschiedene Frequenzbänder zeitlich unterschiedlich bewertet, bspw. bei einem Formanten von /a/ beim Übergang zu /s/. Man

spricht von zeitlicher (engl.: temporal) Modulation. Diese Modulation kann sowohl für das gesamte Frequenzband der Sprache als auch in Subbändern betrachtet werden. Die Zerlegung des Sprachsignals in einzelne Frequenzbänder (z. B. mittels einer Filterbank) führt dazu, dass in jedem Frequenzband (Filterbankkanal c) ein zeitlich schwankender Energieverlauf entsteht. Dieser kann in einer groben Modellvorstellung als Amplitudenmodulation (AM) eines Trägersignals (z. B. eines Rauschens[1] $n_{s,c}(t)$)

$$s(t) = \sum_{c=1}^{C} \mathrm{e}_{s,c}(t) \cdot n_{s,c}(t) \tag{4.6}$$

beschrieben werden, wobei C die Anzahl der Kanäle und $\mathrm{e}_{s,c}(t)$ die zeitliche Einhüllende (engl.: TME – Temporal Modulation Envelope) oder Amplitudenmodulierende von $n_{s,c}(t)$ im entsprechenden Kanal darstellt.

Modulationsspektrum Um den Anteil bestimmter Modulationsfrequenzen f_m (bzw. Modulationskreisfrequenzen ω_m) an der Modulation zu ermitteln, kann die Fouriertransformierte der zeitlichen Einhüllenden gebildet werden. Es entsteht das Modulationsspektrum $\underline{M}_{s,c}(\omega_m)$ (eines Kanals c des Signals s). Dabei haben sich zwei unterschiedliche Betrachtungsweisen herausgebildet:

1. **Modulationsspektrum aus dem TME** $\mathrm{e}_{s,c}(t)$ – Diese Betrachtungsweise wird vorwiegend von Greenberg u. a. (vgl. folgender Abschnitt) vertreten. Das Modulationsspektrum berechnet sich nach
$$\underline{M}_{s,c}(\omega_m) \quad = \quad \mathcal{F}\left\{\mathrm{e}_{s,c}(t)\right\}, \tag{4.7}$$
womit im Wesentlichen die Theorie verfolgt wird, dass die Einhüllende des Signals durch die Bewegung des Vokaltraktes geformt wird.
2. **Modulationsspektrum aus dem TPE** $\mathrm{e}^2_{s,c}(t)$ – Die zeitliche Einhüllende der Leistungsfunktion (engl.: TPE – Temporal Power Envelope) des Signals entspricht $\mathrm{e}^2_{s,c}(t)$ und wird als die Fluktuation von Energie interpretiert, die durch die Bewegung des Vokaltrakts hervorgerufen wird. Bildet man aufgrund von $\mathrm{e}^2_{s,c}(t)$ das Modulationsspektrum, ergibt sich
$$\underline{M}_{s,c}(\omega_m) \quad = \quad \mathcal{F}\left\{\mathrm{e}^2_{s,c}(t)\right\}. \tag{4.8}$$
Diese Betrachtungsweise wird insbesondere in der Theorie der MTF (Houtgast, Steeneken, Schröder und weitere, vgl. Abschnitt 4.3.2) und den daraus motivierten Ansätzen wie RASTA-Filterung (Hermansky et al., vgl. Abschnitt 5.5.2) bzw. MTF-basierte Enthallung (z. B. Unoki et al., vgl. Abschnitt 5.5.3) benutzt und soll auch in dieser Arbeit verwendet werden.

[1] Es ist anzumerken, dass für Sprache selbstverständlich trotzdem das Quelle-Filter-Modell gilt. In stimmhaften Regionen kann allerdings für das Modell der zeitlichen Einhüllenden auch ein Rauschen als Träger dienen. Es wäre auch ein periodischer Träger, wie die stimmhafte Anregungsfunktion, denkbar; für das Modell der zeitlichen Einhüllenden spielt es aber zunächst keine Rolle.

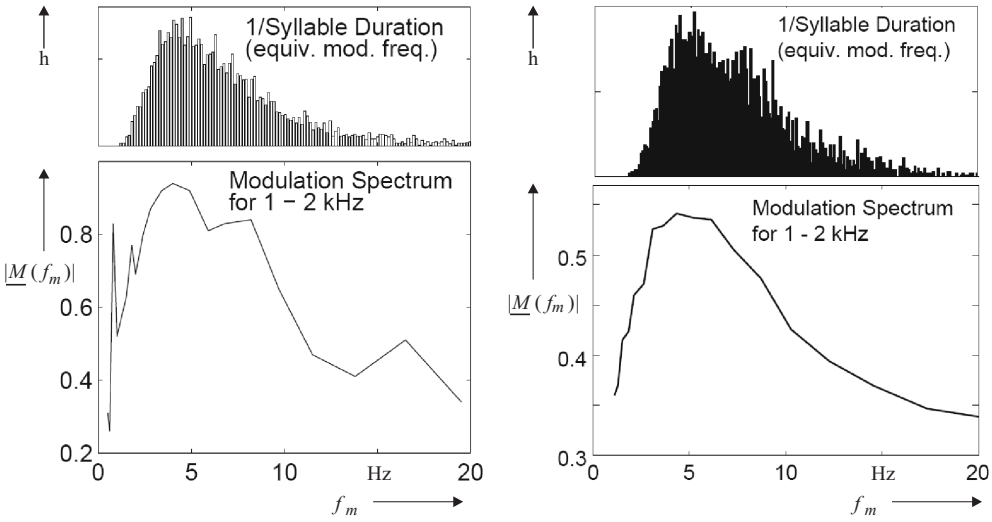

Abbildung 4.2 – Zur Verteilung der Modulationsfrequenzen menschlicher Sprache. Links: Englisch nach Greenberg et al. [GHE96]. Rechts: Japanisch nach Arai et al. [Ara97]. Oben: Histogramm von Silbendauern aus einer signifikanten Stichprobe, auf der Abszisse ist der Kehrwert der Silbendauern aufgetragen, der laut [Gre97, Ara97] mit der Modulationsfrequenz zusammenhängt. Unten: Modulationsspektrum im Frequenzband (1 ... 2) kHz. Es ist zu erkennen, dass die Modulationsinformationen sich zwischen 1 Hz $\leq f_m \leq$ 20 Hz verteilen und ein besonders wichtiger Bereich zwischen 2 Hz $\leq f_m \leq$ 8 Hz liegt. Graphik aus [Ara97] (bzw. nur links Original aus [GHE96]).

Avendano und Hermansky (und weitere) betrachten anstelle der TPEs sogenannte zeitliche Trajektorien von Komponenten (Frequenzen oder Frequenzbändern) aufeinander folgender Leistungsspektren (engl.: Short Term Power Spectrum Trajectory (STPT) [AH96]). Da diese wie bei einem Spektrogramm nur frameweise einen Wert liefern, beschreiben sie eine Annäherung der Abtastung der TPEs (Abtastrate f_s) mit der Framerate (Abtastrate $\frac{1}{FI}$). Von einer exakten Abtastung kann nicht gesprochen werden, da die FFT, die zur Berechnung der Leistungsspektren benutzt wird, durch die Fensterung bereits geringe Abweichungen einbringt. Des Weiteren besitzt die FFT eine zeitliche Unschärfe, wodurch man die genaue zeitliche Position eines Energieereignisses nicht bestimmen kann (zur Unschärferelation z. B. in [Hof98]). Dennoch sind beide, TPEs und STPTs, für die Berechnung des Modulationsspektrums geeignet, da relevante Modulationsfrequenzen, wie im Folgenden gezeigt wird, weit unterhalb der Framerate liegen (Framerate entspricht Überabtastung) und geringe Ungenauigkeiten in Kauf genommen werden können.

Messungen (z. B. in Abbildung 4.2) ergeben, dass bei sauberer Sprache die stärksten Komponenten im Modulationsspektrum zwischen 2 Hz $\leq f_m \leq$ 8 Hz mit ei-

nem Maximum bei 4 Hz liegen, was in etwa mit der mittleren Silbenfrequenz übereinstimmt [Ara97, Gre99, Gre97]. Das entspricht im Wesentlichen den Frequenzen, mit denen die mechanischen Bewegungen des Vokaltrakts beschrieben werden können ($2 \leq f_{Vok.} \leq 12$ Hz [SBMK93]). Unterhalb des Hörspektrums liegend, werden diese Frequenzen nicht als Schall, aber als Energiefluktuationen wahrgenommen, bei deren Frequenzen das Hörzentrum im Gehirn besonders sensibel anspricht ($f_{\text{Hörz.}} < 20$ Hz [SU86, Gre07], maximale Sensibilität des Hörzentrums bei $f_m \approx 4$ Hz [Her97, Gre97]).

Einfluss wichtiger Modulationsfrequenzen auf Sprachverständlichkeit und Spracherkennung Mehrere Wissenschaftler haben das Modulationsspektrum, seinen Einfluss auf die Sprachverständlichkeit sowie bereits auch auf die Spracherkennung untersucht. Als erstes berichtet Dudley 1939 [Dud39], dass tieffrequente Modulationen mit $f_m > 25$ Hz ohne signifikanten Einfluss auf die Sprachverständlichkeit herausgefiltert werden können[2]. Allerdings werden Dudleys Erkenntnisse mit Ausnahme von Zadeh, der den Begriff „variational frequency“ in den 1950er Jahren einführt, und Holmes, der Modulation zur artikulatorischen Synthese in den 1960er Jahren untersucht, Jahrzehnte lang ignoriert [Gre07]. Erst durch die Untersuchungen von Houtgast und Steeneken in den 1970er und 1980er Jahren werden Dudleys Erkenntnisse wieder relevant [HS73, HS80]. Sie demonstrieren, dass eine große Modulationstiefe sowie $f_m < 16$ Hz mit einer guten Sprachverständlichkeit korrelieren und prägen den Begriff Envelope-Spectrum oder Modulation-Spectrum. In den 1980er Jahren werden die Untersuchungen von Houtgast und Steeneken noch mit Skepsis betrachtet, da aus klassischer Sicht das Gehör Frequenzen zwischen ca. 16 und 20000 Hz wahrnehmen kann; wohingegen nun behauptet wird, dass die Modulationsfrequenzen < 16 Hz, die im Bereich des Infraschalls liegen, besonders wichtig für die Sprachverständlichkeit sind [Gre07].

Experimente von Drullman et al. (1994) [DFP94a, DFP94b] widerlegen die Skepsis am Beispiel von Experimenten mit holländischer Sprache, indem sie durch Hoch- bzw. Tiefpassfilterung der Modulierenden nachweisen, dass einige Bereiche im Modulationsspektrum wichtiger für die Verständlichkeit sind als andere. Diese Erkenntnis wird von Arai et al. (1996) [Ara96] auch für das Japanische und von Greenberg (1996) [Gre96] für das Englische bestätigt. Die Experimente von Arai et al. zeigen, dass Modulationsfrequenzen im Bereich von $1 \text{ Hz} \leq f_m \leq 16$ Hz linguistische Informationen enthalten, die für die Sprachverständlichkeit wichtig sind (vgl. Abbildung 4.3 links).

Später werden ähnliche Experimente von Kanedera et al. (1999) [KAHP99] auch für die Spracherkennung (inklusive Variation mehrerer Analysator- und Klassifikatorty-

[2] Dudleys originale Aussage von 1939 in [Dud39]: „... the basic nature of speech as composed of audible sound streams on which the intelligence content is impressed of the true message-bearing waves which, however, by themselves are inaudible“. Übersetzung in die moderne Wissenschaftssprache (2006) nach Atlas in [Atl07]: „Speech and other acoustic signals are actually low bandwidth processes which modulate higher bandwidth carriers“.

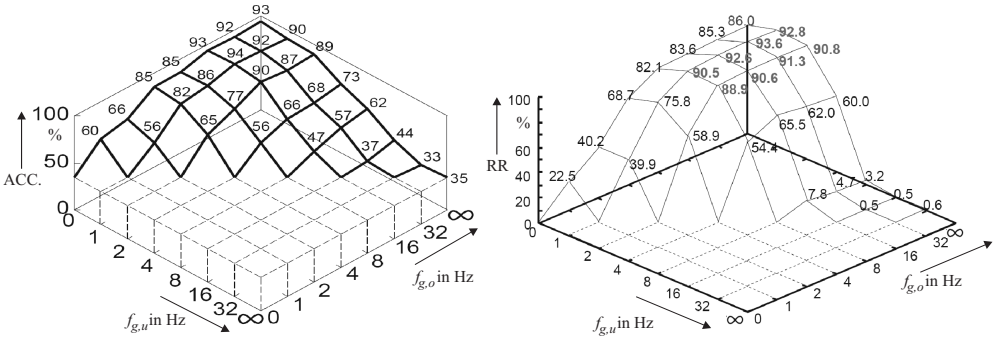

Abbildung 4.3 – Darstellung von Erkennungsexperimenten mit bandpassgefilterten Modulationsfrequenzen. Auf den Abszissen sind die untere $f_{g,u}$ und obere Grenzfrequenz $f_{g,o}$ des Bandpassfilters dargestellt. Links: Menschliche Spracherkennung, Accuracy bei Hörversuchen von japanischen Nonsense-Silben nach Arai [Ara96]. Rechts: Japanische und englische ASR nach Kanedera [KAHP99]. Für die dargestellten Experimente ergibt sich, dass $f_m < 1$ Hz und $f_m > 16$ Hz (links) bzw. $f_m > 32$ Hz (rechts) eher schädlich sind. Graphiken aus [Ara96] und [KAHP99].

pen) durchgeführt (vgl. Abbildung 4.3 rechts). Die Experimente führen zu der Erkenntnis, dass auch bei ASR Modulationsfrequenzen im Bereich von 1 Hz $\leq f_m \leq$ 16 Hz von Bedeutung sind. Im Gegenzug schlussfolgert [KAHP99], dass Modulationsfrequenzen außerhalb dieser Grenzen auf Störungen zurückzuführen sind und demnach die ASR-Performanz verringern.

4.3 Sprachsignal im Raum

4.3.1 Verhalltes Sprachsignal

Die in Gleichung (3.65) dargestellte Faltung beschreibt die Verhallung des ungestörten Signales $s(t)$ durch den Raum. Beachtet man die Dekomposition in Abbildung 3.8 sowie Gleichung (3.67), so fällt auf, dass sich $x(t)$ (wegen des fehlenden Umgebungsrauschens wird $s'(t)$ hier direkt mit $x(t)$ bezeichnet, vgl. Abbildung 3.7) ebenfalls aus einer additiven Überlagerung von Einzelsignalen zusammensetzt, und zwar aus dem direkten und dem diffusen Anteil der RIR $h_{\mathrm{D}}(t)$ bzw. $h_{\mathrm{R}}(t)$. Es entsteht

$$\begin{aligned} x(t) &= (h_{\mathrm{D}} * s)(t) + (h_{\mathrm{R}} * s)(t) \\ &= x_{\mathrm{D}}(t) + x_{\mathrm{R}}(t). \end{aligned} \tag{4.9}$$

Dabei beschreibt $x_{\mathrm{D}}(t)$ das Nutzsignal, das aufgrund der Eigenschaften von $h_{\mathrm{D}}(t)$ unverzerrt ist, aber in der Amplitude vom SMD abhängt. Der Nachhall von $s(t)$ ist

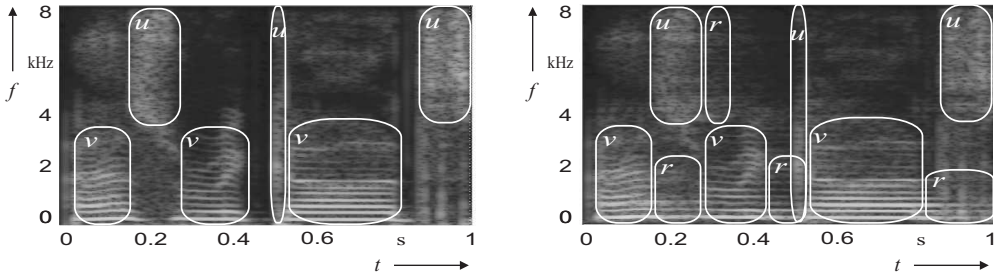

Abbildung 4.4 – Spektrogramm eines Sprachsignals. Links: sauberes Signal $s(t)$. Rechts: verhalltes Signal $x(t)$. Die Signalbestandteile der harmonischen (v) (für engl.: voiced) sowie der nichtharmonischen Komponenten (u) (für engl.: unvoiced) sind sowohl im sauberen als auch im verhallten Signal vorhanden. Im verhallten Signal wird die Hallstörung (r) (für engl.: reverberant) schematisch dargestellt. Die Verhallung erfolgt mit einer RIR eines Wohnzimmers ($T_{60} = 400$ ms, SMD $= 100$ cm). Diese eher schwach hallende Umgebung wurde aus Darstellungsgründen gewählt. Halligere Umgebungen haben einen erheblich stärkeren Effekt, lassen die einzelnen Erscheinungen allerdings nicht so gut graphisch demonstrieren.

das verzerrte Signal $x_{\mathrm{R}}(t)$. Im Gegensatz zu einer stationären additiven Rauschstörung ist die Hallstörung $x_{\mathrm{R}}(t)$ von $s(t)$ abhängig und demnach, analog zu Sprachsignalen, instationär.

Geht man von dem Modell der harmonischen und nichtharmonischen Komponenten im Sprachsignal aus (Gleichung (4.3)), so entsteht mit Gleichung (4.9) das Modell

$$\begin{aligned} x(t) &= (h_{\mathrm{D}} * s_h)(t) + (h_{\mathrm{D}} * s_n)(t) + (h_{\mathrm{R}} * s_h)(t) + (h_{\mathrm{R}} * s_n)(t) \\ &= x_{\mathrm{D},h}(t) + x_{\mathrm{D},n}(t) + x_{\mathrm{R},h}(t) + x_{\mathrm{R},n}(t). \end{aligned} \tag{4.10}$$

Diese Unterteilung wird hier bewusst so dargestellt, um herauszustellen, dass die direkten Komponenten $x_{\mathrm{D},h}(t)$ und $x_{\mathrm{D},n}(t)$ im Signal jeweils in Reinform vorliegen. Sie sind allerdings durch die Hallkomponenten $x_{\mathrm{R},h}(t)$ und $x_{\mathrm{R},n}(t)$ gestört.

Beobachtet man die Signale in Abbildung 4.4, kann durch Überlegungen geschlussfolgert werden:

- Direkte harmonische Komponenten $x_{\mathrm{D},h}(t)$ schauen sehr weit aus dem Spektrum heraus. Demnach können ihre Frequenzmaxima nahezu ungestört gemessen werden. Auf diesem Prinzip beruht bereits der Ansatz HERB von Nakatani et al. [NKM07] (vgl. Abschnitt 5.4.4). Die in Abschnitt 6 eingeführte Methode greift diese Idee ebenfalls auf.
- In Abbildung 4.4 ist zu erkennen, dass der Hall der harmonischen Komponenten $x_{\mathrm{R},h}(t)$ naturgemäß ebenfalls harmonische Strukturen aufweist. Es ist daher an-

zunehmen, dass eine Stimmhaft-Stimmlos-Entscheidung fehlerhaft reagiert, da u. U. in Hallphasen von stimmhaften Lauten eine Grundfrequenz gefunden werden kann. Dieses Verhalten wird durch die Experimente in Abschnitt 7.6.3 auch bestätigt.

- Stimmlose Abschnitte von $s(t)$ sind im tieffrequenten Bereich sehr stark von $x_{\mathrm{R},h}(t)$, herrührend von vorhergehenden stimmhaften Abschnitten, gestört. Wie in Abbildung 4.4 rechts dargestellt wird, sind die tieffrequenten Bereiche in stimmlosen Regionen stark durch einen vorhergehenden hineinhallenden stimmhaften Laut gestört. Durch ihre unterschiedliche Anregung haben stimmhafte Laute tendenziell mehr Energie als stimmlose (vgl. Tabelle 4.2). Das bedeutet, dass die daraus herrührende Hallstörung u. U. mehr Energie besitzt als der Hall des stimmlosen Lautes. Man kann sogar sagen, dass die Energie des Halls eines stimmhaften Lautes zum Zeitpunkt des stimmlosen Lautes teilweise größer ist als die des stimmlosen Lautes selbst. Dabei werden die störenden Frequenzen durch die energiereichen Frequenzbänder der stimmhaften Laute bestimmt. Diese liegen insbesondere bei den Formanten, wobei die Formantfrequenzen für das Deutsche etwa zwischen (200 ... 2900) Hz verteilt sind (vgl. Abbildung 4.1). Diese Theorie der Störung bestimmter Frequenzen, die gerade bei den Formanten liegen, wird durch die Experimente in Abbildung 4.11 (c) und (d) untermauert.
- Nichtharmonische Signalanteile s_n, die vorwiegend aus Rauschen bestehen, ergeben nach dem Verhallen wieder ein Rauschen. Dabei wird $x_{\mathrm{D},n}(t)$ durch $x_{\mathrm{R},n}(t)$ mit Energie angereichert, also verstärkt, aber eventuell auch verfärbt. Aus diesem Grunde werden in [NKM07] die nichtharmonischen Komponenten aus (4.10) zusammengefasst

$$\begin{aligned} x(t) &= (h_{\mathrm{D}} * s_h)(t) + (h_{\mathrm{R}} * s_h)(t) + (h * s_n)(t) \\ &= x_{\mathrm{D},h}(t) + x_{\mathrm{R},h}(t) + x_n(t). \end{aligned} \tag{4.11}$$

  Der Effekt der Verstärkung kann in den Experimenten in Abbildung 4.11 (c) beobachtet werden.

Diese Aussagen, die durch die Experimente in Abschnitt 4.6 noch unterstützt werden, münden letztlich in die Annahmen, die den in dieser Arbeit entwickelten Ansatz HFA in Abschnitt 6 motivieren.

4.3.2 Einfluss des Raumes auf das Modulationsspektrum – Modulationsübertragungsfunktion

Gleichung (4.6) beschreibt Sprache als amplitudenmoduliertes Trägersignal. Nach Gleichung (3.81) kann eine RIR ebenfalls mit einer Einhüllenden und einem Rauschen

modelliert werden. Faltet man beide, entsteht

$$\begin{aligned} x(t) &= (s * h)(t) && (4.12)\\ &= (\mathrm{e}_s n_s * \mathrm{e}_h n_h)(t) && (4.13)\\ &= \mathrm{e}_x(t) n_x(t). && (4.14) \end{aligned}$$

Verschiedene Wissenschaftler verfolgen die Theorie, dass die Faltung der TPEs von $s(t)$ und $h(t)$ (d. h. unabhängig vom Träger $n_s(t)$ bzw. $n_h(t)$) gerade den TPE von $x(t)$ ergibt

$$\begin{aligned} \mathrm{e}_x^2(t) &= (\mathrm{e}_s^2 * \mathrm{e}_h^2)(t) && (4.15)\\ \mathrm{e}_x^2(t) &= \int_0^\infty \mathrm{e}_s^2(t-\tau)\mathrm{e}_h^2(\tau)\mathrm{d}\tau. && (4.16) \end{aligned}$$

Vorraussetzung für diese Aussage ist, dass $n_s(t)$ und $n_h(t)$ voneinander stochastisch unabhängig sind. Der mathematische Beweis ist z. B. im Anhang von [UFSA04] zu finden. Ohne Beweis wird dies bereits in [HSP80] behauptet, allerdings spricht [HSP80] von der Intensität (entspricht der mittleren Leistung des Signals, Gleichung (3.9)) und nicht von der quadrierten Einhüllenden. Unterstellt man jedoch, dass TPEs und mittlere Leistung über einen Proportionalitätsfaktor verknüpft sind, gelten die Ausführungen in [HSP80] auch für TPEs. Abbildung 4.5 beschreibt diesen Zusammenhang vereinfachend für eine einzige sinusförmige Amplitudenmodulierende $\mathrm{e}_s(t)$

$$\begin{aligned} \mathrm{e}_s(t) &= \hat{\mathrm{e}}_s |\cos(\frac{\omega_m}{2}t)| && (4.17)\\ s(t) &= \hat{\mathrm{e}}_s |\cos(\frac{\omega_m}{2}t)| \cdot n_s(t) && (4.18)\\ \mathrm{e}_s^2(t) &= \frac{\hat{\mathrm{e}}_s^2}{2}(1+\cos(\omega_m t)) && (4.19)\\ s^2(t) &= \frac{\hat{\mathrm{e}}_s^2}{2}(1+\cos(\omega_m t)) \cdot n_s^2(t) && (4.20) \end{aligned}$$

mit der Amplitude der Einhüllenden $\hat{\mathrm{e}}_s$. Setzt man (4.19) in (4.16) ein, entsteht (einfacher in komplexer Schreibweise der Kosinusfunktion)

$$\begin{aligned} \mathrm{e}_x^2(t) &= \int_0^\infty \frac{\hat{\mathrm{e}}_s^2}{2}\Big(1+\underbrace{\Re\Big\{e^{j\omega_m(t-\tau)}\Big\}}_{\cos(\omega_m(t-\tau))}\Big)\mathrm{e}_h^2(\tau)\mathrm{d}\tau && (4.21)\\ &= \frac{\hat{\mathrm{e}}_s^2}{2}\Big(\underbrace{\int_0^\infty \mathrm{e}_h^2(\tau)\mathrm{d}\tau}_{W_{\mathrm{e}_h}} + \Re\Big\{e^{j\omega_m t}\underbrace{\int_0^\infty \mathrm{e}_h^2(\tau)e^{-j\omega_m\tau}\mathrm{d}\tau}_{\underline{M}_h(\omega_m)\ \ (komplex)}\Big\}\Big) && (4.22) \end{aligned}$$

$$= \frac{W_{\mathrm{e}_h} \cdot \hat{\mathrm{e}}_s^2}{2}\left(1 + \Re\left\{\frac{\underline{M}_h(\omega_m)}{W_{\mathrm{e}_h}} e^{j\omega_m t}\right\}\right) \tag{4.23}$$

$$= \frac{\hat{\mathrm{e}}_x^2}{2}\Big(1 + m(\omega_m)\cos(j\omega_m t)\Big). \tag{4.24}$$

mit

$$\hat{\mathrm{e}}_x^2 = W_{\mathrm{e}_h}\, \hat{\mathrm{e}}_s^2 \tag{4.25}$$

$$\underline{M}_h(\omega_m) = \mathcal{F}\left\{\mathrm{e}_h^2(t)\right\} = \int_0^\infty \mathrm{e}_h^2(t) e^{-j\omega_m t}\mathrm{d}t, \tag{4.26}$$

$$\underline{m}(\omega_m) = \frac{\underline{M}_h(\omega_m)}{W_{\mathrm{e}_h}} = \frac{\int_0^\infty \mathrm{e}_h^2(t) e^{-j\omega_m t}\mathrm{d}t}{\int_0^\infty \mathrm{e}_h^2(t)\mathrm{d}t} \tag{4.27}$$

$$m(\omega_m) = \Big|\underline{m}(\omega_m)\Big|. \tag{4.28}$$

Der Parameter W_{e_h} wird in der Literatur üblicherweise mit a bezeichnet [HSP80]. Dennoch wird hier W_{e_h} gewählt, da er der Energie der TME-Funktion $\mathrm{e}_h(t)$ (bzw. von $h(t)$ bei [HSP80]) entspricht und a hier bereits anderweitig benutzt wird. $\underline{M}_h(\omega_m)$ ist die Fouriertransformierte von $\mathrm{e}_h^2(t)$ und damit für die TPEs die Übertragungsfunktion des Raumes (4.15). Sie verändert Amplitude und Modulationstiefe $m(\omega_m)$ der TPEs (vgl. (4.24)) bei den verschiedenen Modulationsfrequenzen ω_m. $m(\omega_m)$ bezieht sich durch die Normierung (4.27) im Gegensatz zu $\underline{M}_h(\omega_m)$ nur auf den Unterschied in der Modulationstiefe zwischen $\mathrm{e}_s^2(t)$ und $\mathrm{e}_x^2(t)$ und wird deshalb als Modulationsübertragungsfunktion[3] (engl.: Modulation Transfer Funktion – MTF [HS73]) des Raumes bezeichnet. Durch (4.28) wird aus der sogenannten komplexen MTF (CMTF) $\underline{m}(\omega_m)$ die reelle MTF, mit der normalerweise gearbeitet wird[4]. In (4.19) beträgt die Modulationstiefe 1, d. h., die Maxima und Minima von $\mathrm{e}_s^2(t)$ erreichen $\hat{\mathrm{e}}_s^2$ bzw. 0. Die Abbildungen 4.5 (h) und (i) beschreiben, wie die Modulationstiefe durch den Raum abflacht ($m(\omega_m) < 1$), d. h., die Maxima und Minima der Modulationsfunktion (cos)

[3] Die Notation des kleinen $m(\omega_m)$ der MTF, bzw. des Modulationsindex, stammt aus [HS73, HSP80] bzw. [Sch81]. Aus historischen Gründen wird, obwohl es sich um eine Übertragungsfunktion handelt, für die MTF ein kleiner Buchstabe gewählt. Dennoch benutzen bspw. Unoki [USFA04], Habets [Hab06] oder Langhans und Strube [LS82] für die komplexe MTF ein großes $\underline{M}(\omega_m)$. Für die reelle MTF wird in den meisten Veröffentlichungen zu diesem Thema, auch in [USFA04], ein kleines $m(\omega_m)$ gewählt. In dieser Arbeit beschreibt das große $\underline{M}$ das Modulationsspektrum (Unoki wählt ein E in [USFA04]).

[4] Die Betragsbildung in (4.28) bzw. beim Übergang von (4.23) nach (4.24) ist eine Annäherung, die die Phase in $\underline{M}_h(\omega_m)$ nicht berücksichtigt (so auch in [Sch81]). Diese Ungenauigkeit spiegelt sich zwar im TPE als Phasenverschiebung wider, für die Modulationstiefe ist diese jedoch nicht von Bedeutung.

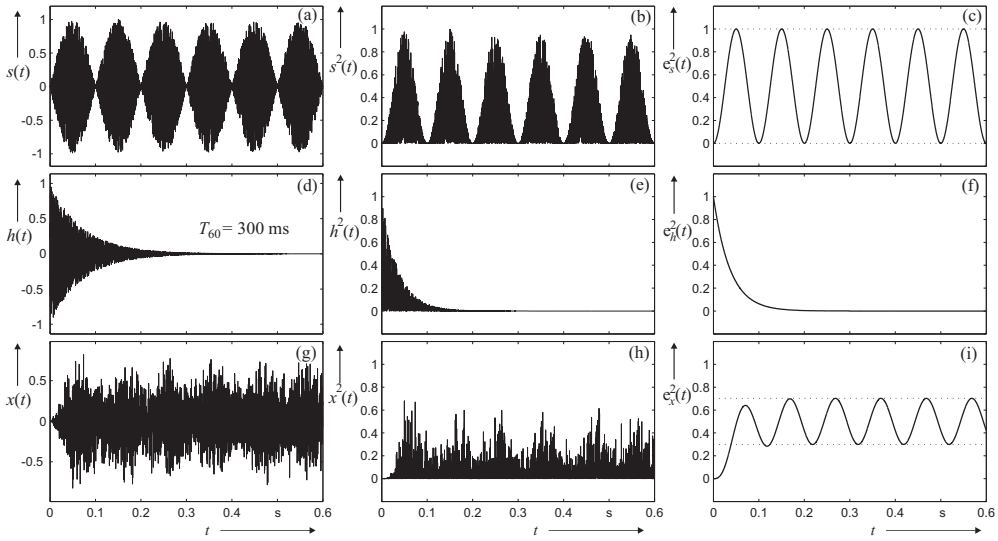

Abbildung 4.5 – Darstellung der Verringerung der Modulationstiefe eines Kosinus-modulierten Rauschsignals $s(t)$ bei Verhallung. (a) unverhalltes Signal $s(t)$ (berechnet nach (4.18)), (b) unverhalltes Leistungssignal $s^2(t)$ (berechnet nach (4.20)), (c) unverhallter TPE $\mathrm{e}_s^2(t)$ von $s(t)$ (berechnet nach (4.19)), (d) RIR $h(t)$ im Fernfeld mit $T_{60} = 500$ ms, (e) Leistungsfunktion der RIR $h^2(t)$, (f) TPE der RIR $\mathrm{e}_h^2(t)$, (g) verhalltes Signal $x(t)$ (berechnet nach (4.12)), (h) verhalltes Leistungssignal $x^2(t)$ und (i) TPE $\mathrm{e}_x^2(t)$ des verhallten Signals $x(t)$ (berechnet nach (4.15)). Die Verringerung der Modulationstiefe ist in der normierten Darstellung (c) vs. (i) ablesbar. (a) – (f) sind auf das jeweilige Maximum normiert. (g) – (i) sind auf $\hat{\mathrm{e}}_x$ bzw. $\hat{\mathrm{e}}_x^2$ normiert, vgl. (4.25). Die Graphiken sind in Anlehnung an [UFSA04] erstellt. Teilabbildungen (c) und (f) sind ähnlich auch in [Kut00] bzw. (b) und (e) bereits in [HS73] dargestellt.

erreichen am Empfängerort nicht mehr das Maximum $\hat{\mathrm{e}}_x^2$ und auch nicht mehr die 0. Man kann sich leicht vorstellen, dass die Modulationstiefe m von ω_m abhängig ist. Schnelle Schwankungen (ω_m groß) sind bei gleicher Nachhallzeit stärker gestört als langsame (ω_m klein). Dies entspricht einem Tiefpassverhalten[5] (vgl. Abbildung 4.6). In der Literatur wird Gleichung (4.27) normalerweise nicht wie hier mit den TPEs $\mathrm{e}_h^2(t)$, sondern mit den Leistungsfunktionen $h^2(t)$ geschrieben. Wie bereits erwähnt darf angenommen werden, dass der TPE und die Funktion des Erwartungswertes der Leistung (vgl. z. B. (3.80)) über einen Proportionalitätsfaktor a_h zusammenhängen

[5] Diese Aussage gilt für Räume. In der allgemeinen Theorie der MTF, die nicht aus der Akustik stammt, sind aber auch andere Übertragungssysteme vorstellbar, deren Übertragungsverhalten nicht notwendigerweise einen Tiefpass darstellen muss.

$$\mathrm{e}_h^2 \quad = \quad a_h \cdot w_h^2(t). \tag{4.29}$$

Betrachtet man (4.15) im Frequenzbereich, entsteht aus der Faltung die Multiplikation

$$\underline{M}_x(\omega_m) \quad = \quad \underline{M}_h(\omega_m) \cdot \underline{M}_s(\omega_m). \tag{4.30}$$

Gleichung (4.30) ist eine einfache Möglichkeit, $\underline{M}_x(\omega_m)$ und somit $\mathrm{e}_x^2(t)$ zu berechnen. Für $\underline{M}_h(\omega_m)$ ist sowohl eine Lösung für das Nahfeld als auch für das Fernfeld zu bestimmen.

Fernfeldapproximation Befindet sich der Empfänger im Fernfeld, so ergibt sich das Modell der RIR nach Gleichung (3.88) bzw. (3.89) mit der Leistungs- bzw. Schallenergiedichtefunktion. Mit (4.29) ergibt sich daraus

$$\mathrm{e}_h^2(t) = \mathrm{e}_{h,\mathrm{R}}^2(t) = a_h \frac{1}{r_\mathrm{R}^2} \frac{13,8}{T_{60}} \cdot e^{-\frac{13,8}{T_{60}}t}. \tag{4.31}$$

Der Index h,R deutet darauf hin, dass die RIR im Fernfeld von der Hallphase dominiert wird. Der Parameter W_{e_h} aus den Gleichungen (4.22) bis (4.27) ergibt sich zu

$$W_{\mathrm{e}_h} \quad = \quad \int_0^\infty \mathrm{e}_h^2(t)\mathrm{d}t \tag{4.32}$$

$$= \quad \int_0^\infty a_h \frac{1}{r_\mathrm{R}^2} \frac{13,8}{T_{60}} \cdot e^{-\frac{13,8}{T_{60}}t}\mathrm{d}t \quad = \quad -a_h \frac{1}{r_\mathrm{R}^2} \left[e^{-\frac{13,8}{T_{60}}t}\right]_0^\infty \tag{4.33}$$

$$= \quad a_h \frac{1}{r_\mathrm{R}^2}. \tag{4.34}$$

Die Fouriertransformierte des TPEs ergibt sich demnach zu

$$\underline{M}_h(\omega_m) \quad = \quad \underline{M}_{h,\mathrm{R}}(\omega_m) \quad = \quad \mathcal{F}\left\{\mathrm{e}_{h,\mathrm{R}}^2(t)\right\} = a_h \frac{1}{r_\mathrm{R}^2} \frac{13,8}{T_{60}} \mathcal{F}\left\{e^{-\frac{2 \cdot 6,9}{T_{60}}t}\right\} \tag{4.35}$$

$$= \quad a_h \frac{1}{r_\mathrm{R}^2} \frac{13,8}{T_{60}} \int_0^\infty e^{-\frac{13,8}{T_{60}}t} \cdot e^{-j\omega_m t}\mathrm{d}t \tag{4.36}$$

$$= \quad a_h \frac{1}{r_\mathrm{R}^2} \frac{13,8}{T_{60}} \left[\frac{-1}{\frac{13,8}{T_{60}} + j\omega_m} e^{-(\frac{13,8}{T_{60}} + j\omega_m)t}\right]_0^\infty \tag{4.37}$$

$$= \quad a_h \frac{1}{r_\mathrm{R}^2} \frac{13,8}{T_{60}} \frac{1}{\frac{13,8}{T_{60}} + j\omega_m} \tag{4.38}$$

$$= \quad a_h \frac{1}{r_\mathrm{R}^2} \cdot \frac{1}{1 + j\omega_m \frac{T_{60}}{13,8}} \tag{4.39}$$

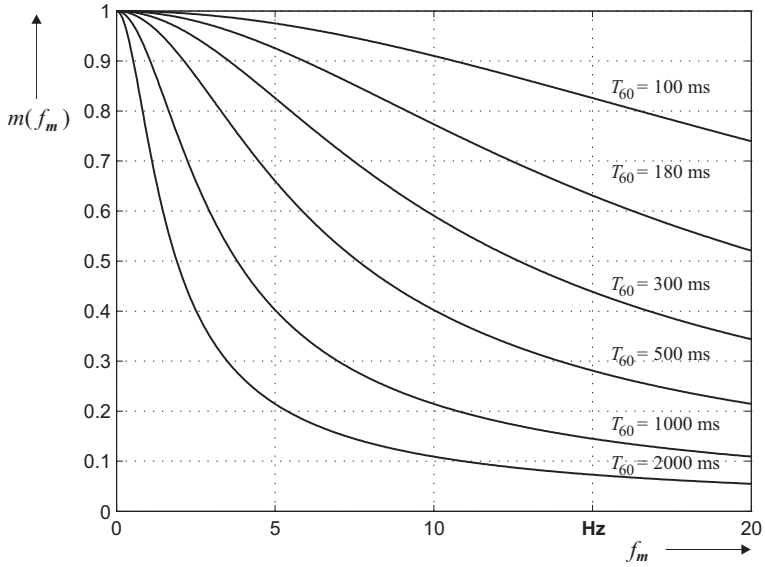

Abbildung 4.6 – MTF im Fernfeld für verschiedene Nachhallzeiten. Anstelle der Modulationskreisfrequenz ω_m wird die zugehörige Modulationsfrequenz $f_m = \frac{\omega_m}{2\pi}$ benutzt.

Die komplexe MTF ergibt sich mit (4.27) zu

$$\underline{m}(\omega_m) = \underline{m}_{\mathrm{R}}(\omega_m) = \frac{1}{1 + j\omega_m \frac{T_{60}}{13{,}8}}, \tag{4.40}$$

und nach Betragsbildung entsteht die reelle MTF einer RIR im Fernfeld

$$m(\omega_m) = |\underline{m}_{\mathrm{R}}(\omega_m)| = \sqrt{\frac{1}{1 + \left(\omega_m \frac{T_{60}}{13{,}8}\right)^2}}. \tag{4.41}$$

Diese beiden Gleichungen leitet Schröder in [Sch81] aus der RIR und nicht aus den TPEs mit[6]

$$\underline{m}(\omega_m) = \underline{m}_{\mathrm{R}}(\omega_m) = \frac{\mathcal{F}\left\{h^2(t)\right\}}{W_h} = \frac{\int_0^\infty h^2(t) \cdot e^{-j\omega_m t}\mathrm{d}t}{\int_0^\infty h^2(t)\mathrm{d}t} \tag{4.42}$$

[6] Anmerkung: An vielen Literaturstellen wird in dieser Formel das Negationszeichen im Exponenten der Fouriertransformation im Zähler weggelassen, z. B. in [Kut00, KM04], [Hab06] oder [UFSA04]. Dabei handelt es sich offensichtlich um eine weitergetragene Ungenauigkeit. Genaugenommen wird dadurch die inverse Fouriertransformation $\mathcal{F}^{-1}$ gebildet, was in (4.40) zu einer Veränderung der Vorzeichen im Nenner führt. Bei der Betragsbildung in (4.41) verschwindet allerdings dieser Fehler, sodass sich das Weglassen des Negationszeichens praktisch nicht auswirkt.

ab. Es wird ausdrücklich betont, dass dieser Ausdruck nur für die Fernfeldapproximation einer RIR zutrifft und demnach nicht für das Nahfeld gilt. Dieser Sachverhalt ist wichtig für später eingeführte Enthallungsalgorithmen. Abbildung 4.6 zeigt die MTF im Fernfeld für verschiedene T_{60}. Das Tiefpassverhalten des Raumes bezüglich der Modulation ist deutlich zu erkennen. Um die MTF eines Raumes im Fernfeld zu schätzen, ist es lediglich erforderlich, T_{60} zu kennen.

Nahfeldapproximation Im Nahfeld wird zusätzlich zur Hallphase noch der direkte Signalanteil benötigt. Die RIR kann also energetisch nach den Modellen in den Gleichungen (3.91) und (3.92) beschrieben werden. Der TPE für die Hallphase kann wie in (4.31) bestimmt werden. Nach (3.78) ergibt sich der TPE der Direktschallphase zu

$$\mathrm{e}_{h,\mathrm{D}}(t) = a_h \frac{1}{r^2}\delta^2(t). \tag{4.43}$$

a_h begründet sich über die Energiebeziehung wie in (4.31). Der TPE der vollständigen RIR ergibt sich zu

$$\mathrm{e}_h(t) = \begin{cases} a_h \left(\frac{1}{r^2}\delta^2(t) + \frac{1}{r_\mathrm{R}^2} \frac{13{,}8}{T_{60}} \cdot e^{-\frac{13{,}8}{T_{60}} t} \right) & ; \quad t \geq 0 \\ 0 & ; \quad t < 0. \end{cases} \tag{4.44}$$

Durch Anwendung von (4.8) entsteht

$$\underline{M}_h(\omega_m) = \mathcal{F}\left\{\mathrm{e}_h^2(t)\right\} = \mathcal{F}\left\{\mathrm{e}_{h,\mathrm{D}}^2(t)\right\} + \mathcal{F}\left\{\mathrm{e}_{h,\mathrm{R}}^2(t)\right\} \tag{4.45}$$

$$= a_h \left[\frac{1}{r^2} + \frac{1}{r_\mathrm{R}^2} \cdot \frac{1}{1 + j\omega_m \frac{T_{60}}{13{,}8}} \right]. \tag{4.46}$$

Die komplexe MTF kann nach (4.27)

$$\underline{m}(\omega_m) = \frac{\frac{1}{r^2} + \frac{1}{r_\mathrm{R}^2} \cdot \frac{1}{1 + j\omega_m \frac{T_{60}}{13{,}8}}}{\frac{1}{r^2} + \frac{1}{r_\mathrm{R}^2}} \tag{4.47}$$

und die MTF durch Betragsbildung nach (4.28)

$$m(\omega_m) = \left| \frac{r^2 r_\mathrm{R}^2}{r^2 + r_\mathrm{R}^2} \left(\frac{1}{r^2} + \frac{1}{r_\mathrm{R}^2} \cdot \frac{1}{1 + j\omega_m \frac{T_{60}}{13{,}8}} \right) \right| \tag{4.48}$$

berechnet werden. Es wird sofort ersichtlich, dass die mathematische Struktur der MTF im Nahfeld weitaus komplizierter als die im Fernfeld ist. Um die MTF eines Raumes im Nahfeld zu schätzen, ist als weitere Schwierigkeit außerdem nicht nur die Kenntnis von T_{60}, sondern auch von r_R und r von Nöten.

4.4 Subjektive und objektive Maße zur Bestimmung der Hallstörung

In diesem Abschnitt wird versucht, ein geeignetes Maß für den Grad der Gestörtheit von Sprache durch Hall zu bestimmen, das für die Spracherkennung angewendet werden kann. Nach einer Einleitung folgen einige ausgewählte Maße, die sich mit der Störung durch Hall befassen.

4.4.1 Einführung

In der Raumakustik sind seit den 30er Jahren des vergangenen Jahrhunderts traditionelle Maße entwickelt worden, die die Hörsamkeit[7] in Räumen quantifizieren. Die Motivation war, Eigenschaften von Räumen bzw. Konzertsälen insbesondere für Sprachverständlichkeit oder Musikdarbietungen zu messen und zu optimieren. Dabei wird zwischen subjektiven und objektiven Maßen [SR84] unterschieden. Subjektive Maße für Sprache sind z. B. die Silben-, Wort- oder Satzverständlichkeit[8] oder ein Mean Opinion Score[9] (MOS). Sie werden durch Hörversuche bei bestimmten akustischen Umgebungen (hier Raumhall) ermittelt [SR84]. Objektive Maße können messtechnisch ermittelt werden, oft aus einer aufgenommenen RIR. Bis zum Ende der 60er Jahre waren dies im Wesentlichen die Deutlichkeit, Hallabstand, Hallmaß, Anfangsnachhallzeit, Rise-Time und Schwerpunktzeit (zusammenfassend definiert in [Kue69]; auf die letzten drei wird hier nicht weiter eingegangen). Bis zum Ende der 70er Jahre kommen Weitere hinzu. Einen umfassenden Überblick aus dem Jahre 1984 findet man in [SR84]; hier werden ca. 30 Maße definiert, die aus der Raumimpulsantwort abgelesen werden können. Neuer, aber etwas weniger umfassend ist [KM04].

Die meisten Maße, die für die Hörsamkeit von Räumen wichtig sind, unterscheiden zwischen:

- **frühen Reflexionen** – Frühe Reflexionen verändern zwar das originale Signal, werden aber vom Hörer nicht als separate Töne und damit auch nicht als Störung wahrgenommen. Die Veränderung äußert sich für den Menschen in einer Verfärbung des Signals (engl.: Colouration) [Kut00], da frühe Reflexionen in der RIR teilweise bereits kein weißes Rauschen mehr darstellen. Der Grund dafür liegt in der frequenzabhängigen Reflexion verschiedener Materialien (vgl. Tabelle 3.2). Direktschall wird durch frühe Reflexionen energiereicher/lauter wahrgenommen, da sie das Originalsignal, welches mit steigendem SMD im-

[7] Raumakustische Gesamtqualität [SR84]; Oberbegriff für die akustischen Wirkungen eines Raumes auf einen Hörer.

[8] Prozentsatz der bei einem Hörtest richtig verstandenen Silben, Wörtern oder Sätzen [SR84].

[9] Der MOS-Wert ist ein Maß, welches die subjektiv wahrgenommene Qualität von Sprache durch eine Gruppe von Versuchspersonen beurteilt. Eine Empfehlung zur Ermittlung des MOS-Wertes findet man in [ITU96].

mer schwächer wird (Gleichung (3.27)), unterstützen. Frühe Reflexionen werden deshalb als nützliche Reflexionen bezeichnet. In Klassenräumen, Hör- oder Konzertsälen etc. spielen nützliche Reflexionen eine wichtige Rolle bei der akustischen Bauplanung.

- **späten Reflexionen** – Späte Reflexionen werden vom Menschen als Störung wahrgenommen. Sie addieren die diffuse Schallenergie auf das Nutzsignal. Die späten, diffusen Reflexionen führen dazu, dass die Störung als Verschleifung bzw. Verhallung des originalen Signals wahrgenommen wird. Diese ist vom Menschen hörbar und wirkt störend auf die Verständlichkeit von Sprache bzw. die Durchsichtigkeit[10] von Musik.
- **diskreten Reflexionen** – Wenn eine deutlich separate Reflexion wahrgenommen wird (Wiederholung), spricht man von Echos[11,12]. In kleinen Räumen sind solche starken diskreten Reflexionen, wie bereits erwähnt, normalerweise nicht vorhanden. Deshalb spielen sie in dieser Arbeit keine Rolle.

Der Übergang von nützlichen und störenden Reflexionen ist fließend und demnach nicht genau festzulegen [KM04]. Haas et. al. ermitteln durch Experimente zur Silbenverständlichkeit als Funktion von unterschiedlich stark verzögerten Reflexionen, dass eine sogenannte kritische Verzögerungszeit t_k existiert, ab welcher die Reflexionen beginnen, die Sprachverständlichkeit zu stören. t_k liegt bei Haas etwa zwischen 50 und 100 ms [Haa51] (zitiert in [Kut00] und weiteren).

Zur Einführung einiger Maße sollen im Vorfeld folgende Energien definiert werden: In einer RIR mit der Gesamtenergie

$$W_h = \int_0^\infty h^2(t)\mathrm{d}t \tag{4.49}$$

kann man die nützliche

$$W_{h_{\mathrm{nutz},t_k}} = \int_0^{t_k} h^2(t)\mathrm{d}t \tag{4.50}$$

und die störende Signalenergie

$$W_{h_{\mathrm{stör},t_k}} = \int_{t_k}^\infty h^2(t)\mathrm{d}t = W_h - W_{h_{\mathrm{nutz},t_k}} \tag{4.51}$$

[10] Möglichkeit, aufeinanderfolgende Töne oder Klänge der Musik, die musikalisch bedeutungsvoll sind, deutlich zu erkennen bzw. zu unterscheiden [SR84].

[11] Dafür muss eine bestimmte Laufzeit und Rückwurfstärke eingehalten werden, vgl. Echokriterium in [SR84].

[12] Begriff nicht zu verwechseln mit der Definition bei Anwendung der Acoustic-Echo-Cancellation (AEC) [Pet01]. Bei der AEC wird als akustisches Echo die Rückkopplung durch einen Lautsprecher bei z. B. einer Freisprecheinrichtung bezeichnet. Dabei wird kein Echokriterium etc. verlangt. Vielmehr wird die Summe aller Reflexionen als akustisches Echo bezeichnet.

der RIR messen. Weiterhin sollen noch die Direktschallenergie

$$W_{h_\mathrm{D}} = \int_0^{t_\mathrm{Impuls}} h^2(t)\mathrm{d}t \tag{4.52}$$

(vgl. auch (3.72)), die Energie aller Reflexionen (hier vereinfacht Hallenergie)

$$W_{h_\mathrm{R}} = \int_{t_\mathrm{Impuls}}^{\infty} h^2(t)\mathrm{d}t = W_h - W_{h_\mathrm{D}} \tag{4.53}$$

(vgl. auch (3.83)) sowie die Energie der diskreten Anfangsreflexionen

$$W_{h_{\mathrm{R},i}} = \sum_i \int_{t_{i,\mathrm{start}}}^{t_{i,\mathrm{stop}}} h^2(t)\mathrm{d}t \tag{4.54}$$

definiert werden. $t_{i,\mathrm{start}}$ und $t_{i,\mathrm{stop}}$ beschreiben Anfangs- und Endzeitpunkt der i-ten diskreten Reflexion. t_Impuls wird in [SL74] mit maximal 5 ms beschrieben (ohne Laufzeitverzögerung zwischen Quelle und Empfänger).

4.4.2 Schalleindruck

Der Schalleindruck wird 1935 von Aigner und Strutt vorgestellt [AS35]; die englische Originalbezeichnung lautet Sound Impression. Er wird als erste Größe vorgeschlagen, die auf einem akustischen Energieverhältnis mit dem Prinzip von frühen und späten Reflexionen beruht und ist deshalb der Vollständigkeit halber angegeben. Q wird aus dem Energieverhältnis

$$Q = \frac{W_\mathrm{D} + W_\mathrm{R,früh}}{W_\mathrm{R,spät} + W_\mathrm{noise}} \tag{4.55}$$

der Signalanteile des Sprach- oder Musiksignals (nicht zu verwechseln mit den oben definierten, aus der RIR berechneten Energien) berechnet. Es gilt

$$W_{\ldots} = \int_0^T x_{\ldots}^2(t)\mathrm{d}t \tag{4.56}$$

für die Energie eines bestimmten Signalbestandteils $x_{\ldots}(t)$. Die einzelnen Bestandteile in (4.55) sind das direkte Signal $x_\mathrm{D}(t) = (h_\mathrm{D} * s)(t)$, frühe Reflexionen $x_\mathrm{R,früh}(t) = (h_\mathrm{R,früh} * s)(t)$, späte Reflexionen $x_\mathrm{R,spät}(t) = (h_\mathrm{R,spät} * s)(t)$ sowie Rauschen $x_\mathrm{noise}(t) = n(t)$. Die Grenze zwischen frühen und späten Reflexionen wird in [AS35] mit $\frac{1}{16}$ s angegeben. $Q = 1$ gilt bei [AS35] als befriedigender Schalleindruck. Obwohl Q noch keine ausreichende Lösung ist, bildet die Struktur von (4.55) bereits die Basis für später entwickelte Maße (s. u.).

4.4.3 Deutlichkeit, Deutlichkeitsgrad

Die Deutlichkeit D_{50} wird Anfang der 50er Jahre von Thiele [Thi53] eingeführt. Sie ist ein objektives Maß zur Beschreibung der Sprachverständlichkeit in Räumen. In [SR84] wird der Begriff Deutlichkeitsgrad benutzt.

$$D_{50} = \frac{\int_0^{t_k=50\ \mathrm{ms}} h^2(t)\mathrm{d}t}{\int_0^{\infty} h^2(t)\mathrm{d}t} 100\ \% = \frac{W_{h_{\mathrm{nutz,50\ ms}}}}{W_h} 100\ \% = \frac{W_{h_{\mathrm{nutz,50\ ms}}}}{W_{h_{\mathrm{nutz,50\ ms}}} + W_{h_{\mathrm{stör,50\ ms}}}} 100\ \% \tag{4.57}$$

Verschiedene unabhängige Studien ([Bor53] (zitiert in [Kut00]) bzw. [Nie56, Jan68] (zitiert in [SR84])) untersuchen die Relation zwischen D_{50} und der Silbenverständlichkeit und kommen dabei jeweils auf das Ergebnis, dass bei $D_{50} > 50\ \%$ eine Silbenverständlichkeit von $> 90\ \%$ zu erwarten ist.

4.4.4 Nutz-Stör-Verhältnis

Lochner und Burger entwickeln bereits 1961 ein Maß, um das Verhältnis von Nutz- zu Störschall für die Sprachübertragung zu beschreiben [LB61]. Die Notation η erinnert an die übliche Notation für den Wirkungsgrad

$$\eta = 10\lg \frac{\int_0^{t_k=95\ \mathrm{ms}} h^2(t)\mathrm{d}t}{\int_{t_k=95\ \mathrm{ms}}^{\infty} h^2(t)\mathrm{d}t}\ \mathrm{dB} = 10\lg \frac{W_{h_{\mathrm{nutz,95\ ms}}}}{W_{h_{\mathrm{stör,95\ ms}}}}\ \mathrm{dB}. \tag{4.58}$$

t_k wird mit 95 ms definiert, was, wie im Folgenden dargestellt ist, noch als relativ hoch angesehen werden kann. Die Silbenverständlichkeit wird bei $\eta = (0; -5)$ dB mit ca. (95; 80) % angegeben [LB61] (zitiert in [SR84]).

4.4.5 Hallmaß

Das Hallmaß R wird 1965 von Beranek et al. [BS65] eingeführt. Es gibt an, wie hallig ein Raum wahrgenommen wird (ursprünglich zur Wahrnehmbarkeit der durch den Raum erzeugten Abklingvorgänge bei Musik)

$$R = 10\lg \frac{\int_{t_k=50\ \mathrm{ms}}^{\infty} h^2(t)\mathrm{d}t}{\int_0^{t_k=50\ \mathrm{ms}} h^2(t)\mathrm{d}t}\ \mathrm{dB} = 10\lg \frac{W_{h_{\mathrm{stör,50\ ms}}}}{W_{h_{\mathrm{nutz,50\ ms}}}}\ \mathrm{dB} = 10\lg \frac{W_{h_{\mathrm{stör,50\ ms}}}}{W_{h_{\mathrm{D}}} + W_{h_{\mathrm{R},i}}}\ \mathrm{dB}. \tag{4.59}$$

Im letzten Term von (4.59) wird bereits die in [RSLA74] vorgeschlagene Substitution $W_{\mathrm{nutz,50}} = W_{h_{\mathrm{D}}} + W_{h_{\mathrm{R},i}}$ vorgenommen. Dies impliziert zwei Näherungen. Zum einen wird angenommen, dass unter 50 ms keine diffusen Reflexionen stattfinden, und

zum zweiten geht [RSLA74] davon aus, dass nach 50 ms keine diskreten Reflexionen mehr auftreten. Diskrete Anfangsreflexionen werden explizit als nützlich bezeichnet. In [SR84] wird die Notation H verwendet; diese steht in [RS66, RSLA74] aber bereits für den Hallabstand (s. u.).

4.4.6 Hallabstand

Der Hallabstand H wird 1966 von Reichardt et al. [RS66] eingeführt und beschreibt eine Art Signal-Stör-Abstand (vgl. auch die Struktur des Nutz-Stör-Verhältnisses nach Lochner und Burger)

$$H_{t_k} = 10\lg\frac{\int_0^{t_k} h^2(t)\mathrm{d}t}{\int_{t_k}^{\infty} h^2(t)\mathrm{d}t}\,\mathrm{dB} = 10\lg\frac{W_{h_{\mathrm{nutz},t_k}}}{W_{h_{\mathrm{stör},t_k}}}\,\mathrm{dB} = 10\lg\frac{W_{h_{\mathrm{D}}}}{W_{h_{\mathrm{stör},t_k}} + W_{h_{\mathrm{R},i}}}\,\mathrm{dB}. \quad (4.60)$$

t_k wird sowohl in [RS66], [Kue69] als auch in [RSLA74] bewusst undefiniert gelassen und als problemspezifisch beschrieben. Im letzten Term von (4.60) sagt [RSLA74] im Unterschied zum Reziproken von (4.59), dass diskrete Anfangsreflexionen bei der Berechnung von H zur Störung gezählt werden. Dennoch kann laut [RSLA74] die Näherung $H_{50} \approx -R$ für $t_k = 50$ ms angenommen werden.

4.4.7 Wirksamer Hallabstand

Hallabstand und Hallmaß sollen zunächst als Maße für den Raumeindruck[13] verwendet werden. Dabei wird in [SL74] festgestellt, dass dies nur bedingt möglich ist. Reichardt stellt nachfolgend (1974) Untersuchungen zu einem wirksamen Hallabstand H_w

$$H_w = 10\lg\frac{W_{h_{\mathrm{nutz},25\ \mathrm{ms}}} + W_{h_{\mathrm{R},i,\mathrm{v}}}}{W_{h_{\mathrm{stör},80\ \mathrm{ms}}} + W_{h_{\mathrm{R},i,\mathrm{s}}}}\,\mathrm{dB} \quad (4.61)$$

an, der die Defizite beider Maße kompensiert. Die Zusatzindizes v (vorn) und s (seitlich) in $W_{h_{\mathrm{R},i}}$ unterteilen die diffusen Anfangsreflexionen, welche bei Verzögerungszeiten zwischen (25 ... 80) ms auftreten. v steht für direktschallwirksame Anfangsreflexionen (von vorn, definiert mit $\phi < 40°$; ϕ ist der Einstrahlwinkel, Blickrichtung des Beobachters entspricht $\phi = 0°$) und s steht für raumschallwirksame Anfangsreflexionen (von der Seite, definiert mit $\phi \geq 40°$). Durch Hörversuche wird nachgewiesen, dass H_w ein geeignetes Maß zur Beschreibung des Raumeindruckes ist [Rei75].

[13] Unter Raumeindruck wird in [SL74] und [RSLA74] verstanden: (i) wie bewusst der aus dem Raum reflektierte und nicht direkt von der Quelle stammende Schall wahrgenommen wird, (ii) das Gefühl des Eingehülltseins in das Klanggeschehen, (iii) die Gewinnung einer Vorstellung der Raumdimensionen aus dem eintreffenden Schall.

4.4.8 Raumeindrucksmaß

Das Raumeindrucksmaß R_w wird im gleichen Jahr ebenfalls als objektives Kriterium für Musik in [Leh74] vorgestellt

$$R_w = 10\,\mathrm{lg}\,\frac{W_{\mathrm{nutz,25\ ms}} + W_{h_{\mathrm{R},i,\mathrm{v}}}}{W_{\mathrm{stör,80\ ms}} + W_{h_{\mathrm{R},i,\mathrm{s}}}}\ \mathrm{dB}. \tag{4.62}$$

R_w beschreibt das reziproke Verhältnis von H_w; es darf $R_w = -H_w$ angenommen werden. Die Notation R_w wird hier anstelle von R in [SR84] wegen der Verwechselung mit dem Hallmaß benutzt. 1979 wird ein weiteres Raumeindrucksmaß R_K nach Schmidt [Sch79] eingeführt.

4.4.9 Klarheitsmaß, Musikklarheitsmaß, Durchsichtigkeit, Sprachklarheitsmaß, Deutlichkeitsmaß

Das Klarheitsmaß wird 1975 von Reichardt et al. vorgestellt [RAS75]. Es bildet das nun vergleichsweise einfache Verhältnis von früher zu später Energie in der RIR nach

$$C_{t_k} = 10\,\mathrm{lg}\,\frac{\int_0^{t_k} h^2(t)\mathrm{d}t}{\int_{t_k}^{\infty} h^2(t)\mathrm{d}t}\ \mathrm{dB} = 10\,\mathrm{lg}\,\frac{W_{h_{\mathrm{nutz},t_k}}}{W_{h_{\mathrm{stör},t_k}}}\ \mathrm{dB}, \tag{4.63}$$

wobei wieder eine Zeit t_k zur Unterscheidung dient. Das Klarheitsmaß wird zunächst für die Durchsichtigkeit von Musik entwickelt. Dabei hat sich $t_k = 80$ ms als günstig erwiesen. Es ist demnach ein Musikklarheitsmaß oder Durchsichtigkeitsmaß C_{80} nach Abdel Alim definiert [Abd73]. In der Musik wirken Reflexionen, die früher als 80 ms eintreffen, als räumlichkeitsanreichernd; das Räumlichkeitsgefühl wird als angenehm empfunden. Spätere Reflexionen wirken als Störung. Bei Sprache ist die Verständlichkeit und nicht das Räumlichkeitsgefühl von Bedeutung. Deshalb ist t_k für Sprache, wie bereits bei Thiele, mit 50 ms kürzer festgelegt. Aufgrund dessen wird das Sprachklarheitsmaß oder Deutlichkeitsmaß C_{50} nach Ahnert definiert [Ahn75]. Wie man leicht sieht, entspricht das Klarheitsmaß zunächst der ursprünglichen Definition des Hallabstandes ($C_{50} = H_{50} = -R$ bzw. $C_{80} = H_{80}$, linker Term in (4.60)). Die Größen η, H, C_{50} und C_{80} (und auch H_w) beschreiben demnach den gleichen Sachverhalt, abgesehen von subjektiven, menschlich wahrgenommenen Unterschieden.

4.4.10 STI – Speech Transmission Index

In Abschnitt 4.3.2 wird bereits das Konzept der MTF vorgestellt. Houtgast und Steeneken berichten 1973 in [HS73], dass aus der MTF auf den Grad der Sprachverständlichkeit geschlossen werden kann. Daraus wird später der Sprachübertragungsindex (engl.: Speech-Transmition-Index) entwickelt [HS80]. Dabei werden in einem komplexeren

Verfahren (übersichtlich zusammengefasst in [SR84]) nach verschiedener Anregung des Raumes mit definiert moduliertem Rauschen die entsprechenden Werte der MTF ermittelt. Aus einer vorgeschriebenen Anzahl definiert gewichteter MTF-Werte wird der STI berechnet. Er korreliert hervorragend mit der Sprachverständlichkeit [SR84] und ist deshalb ein professionell genutztes Maß. Da der STI in seiner Ursprungsvariante sehr zeitaufwendig zu messen ist, wird später eine schnellere Variante, der Rapid STI (RASTI), entwickelt [HS84, DIN03b].

4.4.11 SRR – Signal-Hall-Abstand

Die bisher dargestellten Maße sind durch die Verständlichkeit von Sprache oder Musik motiviert, wo frühe Reflexionen als nützlich angenommen werden. Im Falle der Spracherkennung kann davon zunächst nicht ausgegangen werden. Vorerst werden alle Reflexionen als störend angenommen. Um den Grad der Störung zu ermitteln, ist es bei Störgeräuschen üblich, einen Signal-Rausch-Abstand, den SNR (engl.: Signal-to-Noise Ratio), zu messen. Wie beim SNR lässt sich auch die Hallstörung[14] zum Nutzsignal ins Verhältnis setzen; man spricht vom SRR[15] (engl.: Signal-to-Reverberation Ratio). Unter Beachtung von Gleichung (4.9) kann der SRR mit

$$\mathrm{SRR} = 10\lg\frac{\int_0^\infty x_\mathrm{D}^2(t)\mathrm{d}t}{\int_0^\infty x_\mathrm{R}^2(t)\mathrm{d}t}\ \mathrm{dB} = 10\lg\frac{\int_0^\infty (s*h_\mathrm{D})^2(t)\mathrm{d}t}{\int_0^\infty (s*h_\mathrm{R})^2(t)\mathrm{d}t}\ \mathrm{dB} \tag{4.64}$$

definiert werden. Die übliche Definition des SRR geht jedoch nur von der RIR aus

$$\mathrm{SRR} = 10\lg\frac{\int_0^\infty h_\mathrm{D}^2(t)\mathrm{d}t}{\int_0^\infty h_\mathrm{R}^2(t)\mathrm{d}t}\ \mathrm{dB} = 10\lg\frac{W_{h_\mathrm{D}}}{W_{h_\mathrm{R}}}\ \mathrm{dB} = 10\lg\frac{\int_0^{t_k=t_\mathrm{Impuls}} h^2(t)\mathrm{d}t}{\int_{t_k=t_\mathrm{Impuls}}^\infty h^2(t)\mathrm{d}t}\ \mathrm{dB}, \tag{4.65}$$

so z. B. definiert in [Hab06, SK08] und weiteren. Diese Definition wird in der Signalverarbeitung oft benutzt, wobei allerdings die Zeit t_k eine nicht einheitlich definierte Größe ist. In dieser Arbeit wird für den SRR $t_k = 5$ ms festgelegt. Nach (4.65) berechnet, beschreibt der SRR das Verhältnis der Leistungen von Direkt- und Diffusschallkomponente der RIR, nicht notwendigerweise das Verhältnis SRR des Signales aus Gleichung (4.64). Es kann aber wie folgt gezeigt werden, dass $s(t)$ aus (4.64) eliminiert werden kann, wenn es sich dabei um ein stochastisches Signal handelt, das Realisierung eines

[14] Da die Störung, nicht wie bei einem additiven Rauschen, durch die Faltung von einem Originalsignal mit einem System erzeugt wird, findet man in einigen Veröffentlichungen daher auch den Begriff Convolutional Noise. Er wird aber in dieser Arbeit nicht benutzt.

[15] Der Term SRR (bspw. benutzt in [SK08] und weiteren) ist vorzugsweise das korrekte Pendant zum SNR und wird deshalb in dieser Arbeit verwendet. Jedoch findet man in vielen Publikationen den Begriff Direct-to-Reverberation Ratio – DRR; insbesondere auch in der Literatur, die sich mit den später behandelten Ansätzen zur Eliminierung der Störung befasst, z. B. [NG05, Hab06]. Der Begriff DRR wird in dieser Arbeit zur Beschreibung der akustischen Zusammenhänge benutzt, vgl. Abschnitt 3.2.2.

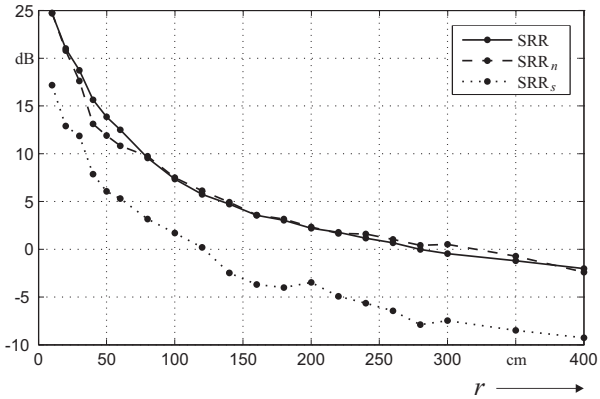

Abbildung 4.7 – Darstellung des SRR für die Messung im SMART-Room in drei Varianten: 1. SRR in seiner üblichen Definition berechnet aus der RIR nach Gleichung (4.65). 2. SRR$_n$ aus einem verhallten Signal berechnet nach Gleichung (4.64), wobei das Testsignal ein Geräusch ist (Index n). 3. SRR$_s$ aus einem verhallten Signal berechnet nach Gleichung (4.64), wobei das Testsignal ein Sprachsignal ist (Index s). Der Vergleich von SRR, SRR$_n$ und SRR$_s$ belegt anschaulich, dass das Eliminieren von $s(t)$ aus (4.64) nur für stochastisch unabhängige Signale (wie das hier verwendete Rauschen) möglich ist. Gleichung (4.65) gilt demnach strenggenommen nicht für Sprachsignale.

stationären ergodischen Prozesses ist. Bei diesen Signalen kann die Leistung über die Erwartungswertbildung berechnet werden

$$\mathrm{SRR} = 10\lg \frac{E\left\{(s * h_\mathrm{D})^2(t)\right\}}{E\left\{(s * h_\mathrm{R})^2(t)\right\}}\ \mathrm{dB}. \tag{4.66}$$

Die Leistung der betrachteten Realisierungen kann mittels der Parseval'schen Beziehung und dem Faltungssatz auch im Frequenzbereich ermittelt werden [Hof98]

$$\mathrm{SRR} = 10\lg \frac{E\left\{S H_\mathrm{D} S^* H_\mathrm{D}^*\right\}}{E\left\{S H_\mathrm{R} S^* H_\mathrm{R}^*\right\}}\ \mathrm{dB}. \tag{4.67}$$

Die Regeln der Erwartungswertbildung enthalten für unkorrelierte Zufallssignale

$$E\left\{\prod X_i(k)\right\} = \prod E\left\{X_i(k)\right\}, \tag{4.68}$$

was auf (4.67) wie folgt

$$\mathrm{SRR} = 10\lg \frac{E\left\{SS^*\right\} E\left\{H_\mathrm{D} H_\mathrm{D}^*\right\}}{E\left\{SS^*\right\} E\left\{H_\mathrm{R} H_\mathrm{R}^*\right\}}\ \mathrm{dB} \tag{4.69}$$

angewendet werden kann, wenn vorausgesetzt wird, dass die entsprechenden Leistungssignale voneinander statistisch unabhängig sind. Die Erwartungswertbildung vollzieht

sich dabei in zeitlicher Richtung, für jedes Frequenzband separat. S kann demnach eliminiert werden und bei Betrachtung im Zeitbereich erhält man Gleichung (4.65).

In Abbildung 3.14 (a) wird die Leistung der im SMART-Room gemessenen RIRs gebildet, wobei $t_k = 5$ ms verwendet wird (die Verzögerungszeit des Direktschalls durch den SMD wird dabei nicht berücksichtigt). Der SRR nach (4.65) ist dementsprechend als Differenz der beiden Graphen abzulesen, er wird in Abbildung 4.7 dargestellt. Zusätzlich ist in dieser Graphik der SRR nach (4.64) für ein Rauschsignal (SRR_n) und für ein Sprachsignal (SRR_s) dargestellt. Die Graphik belegt am Beispiel des Rauschsignals anschaulich, dass $s(t)$, wie oben abgeleitet, aus (4.64) eliminiert werden kann, wenn es sich um ein stochastisches Signal handelt, das Realisierung eines stationären ergodischen Prozesses ist. Für den Fall, dass $s(t)$ ein Sprachsignal ist, kann $s(t)$ nicht eliminiert werden. Durch die gleich bleibende Differenz von etwa 7 dB gegenüber SRR_s kann aus dem SRR dennoch vergleichend auf den Grad der Störung geschlossen werden.

4.4.12 Weitere Maße

In der Dissertation [Hab06] widmet sich ein ganzes Kapitel den Störmaßen. Darin enthalten sind einige der bereits vorgestellten Maße, aber auch weitere, die hier nicht näher ausgeführt werden. Diese sind Segmental SRR ($\mathrm{SRR}_{\mathrm{Seg}}$) [NG05], abgeleitet vom Segmental SNR [QBC88], Log Spectral Distortion (LSD), Bark Spectral Distortion (BSD) [NG05], Reverberation Decay Tail (R_{DT}) [WN06] sowie Perceptual Evaluation of Speech Quality (PESQ) [ITU01].

4.4.13 Störmaß für die Spracherkennung

Es stellt sich nun die Frage, welches Störmaß geeignet ist, das Verhalten der Erkennungsrate bei Störungen zu beschreiben. Für die Sprachverständlichkeit hat sich dabei weltweit das Deutlichkeitsmaß C_{50} durchgesetzt. Ein State-of-the-Art eines Störmaßes für die Spracherkennung existiert nicht bzw. wurde nicht hinreichend untersucht. Es scheint zunächst sinnvoll zu sein, den SRR nach (4.64) dafür zu benutzen. Allerdings ist die Messung aufwendig und nur für bekannte saubere Sprachsignale durchzuführen. Der aus der RIR berechnete SRR nach (4.65) hat den Vorteil, dass er unabhängig vom Sprachsignal ist, d. h., er kann für den realen Einsatz bei einem bekannten Raum vorhergesagt werden. Wie bereits erwähnt, wird dieser auch oft von verschiedenen Wissenschaftlern benutzt. Aus Abbildung 4.7 kann darauf geschlossen werden, dass der SRR nach (4.65) die Störung von Sprache vergleichbar abbildet. In Abbildung 4.8 wird der SRR mit den wichtigsten herkömmlichen Maßen C_{50} und C_{80} verglichen. Die Darstellung oben rechts benutzt wieder die RIRs aus dem SMART-Room. Die Ergebnisse zeigen, dass der SRR im Prinzip geeignet ist. In der linken oberen Graphik ist dem hingegen nicht so. Hier werden die für die Erkennungsexperimente ab Abschnitt

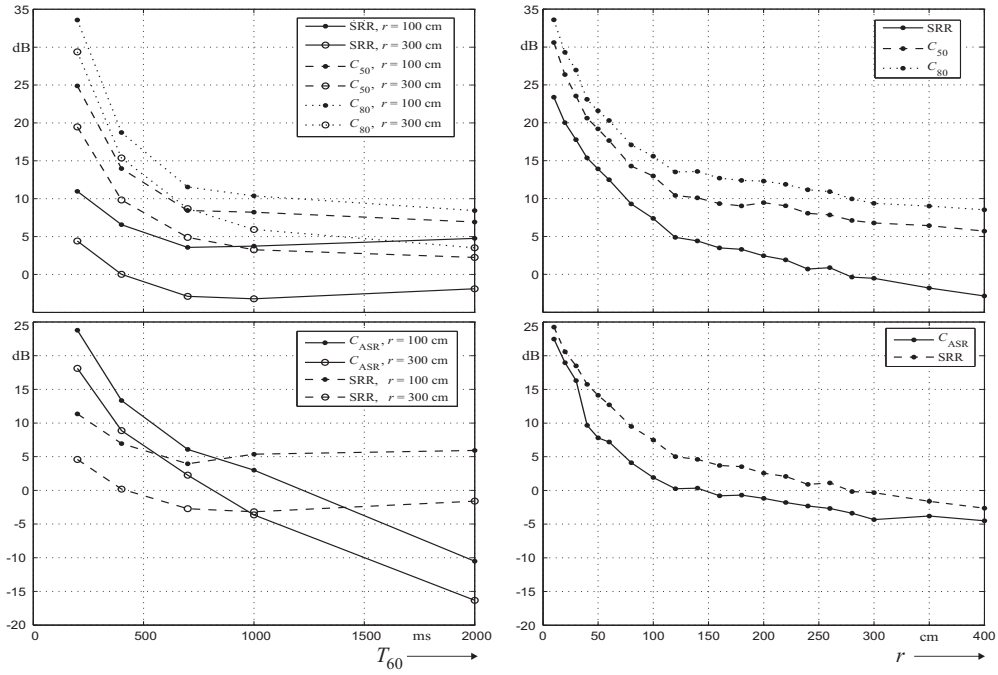

Abbildung 4.8 – Oben: Vergleichende Darstellung von Klarheitsmaß C_{80}, Deutlichkeitsmaß C_{50} und SRR ($t_k = 5$ ms) für die in dieser Arbeit ausgewählten RIRs (ausgewähltes Set für Experimente ab Abbildung 5.3). Unten: Darstellung eines vorgeschlagenen Maßes C_{ASR} anhand des Beispiels aus Gleichung (4.70). Zum Vergleich wird noch der SRR aus der obigen Abbildung hinzugefügt. Links: Abhängigkeit von T_{60}. Rechts: Abhängigkeit vom SMD im SMART-Room.

6 ausgewählten RIRs vermessen. Es ist zu erkennen, dass sich der SRR bei steigendem T_{60} nicht notwendigerweise verringert. Der SRR-Wert für $T_{60} = 2000$ ms (Treppenhaus TU Dresden, Volumen groß) ist bspw. größer als der für $T_{60} = 700$ bzw. 1100 ms (Büro klein und Büro groß). Dies ist bei C_{50} und C_{80} nicht der Fall. Erklärbar ist dieser Effekt zum einen durch das große Volumen, da die diffuse Schallenergiedichte (zumindest bei gleich bleibendem T_{60}) mit steigendem Volumen sinkt (vgl. Gleichung (3.37) bzw. Abbildung 3.2). Die scheinbar korrektere fallende Tendenz bei C_{50} und C_{80} ist wiederum erklärbar durch das Einbeziehen der frühen Reflexionen, die in kleineren Räumen durch die kürzeren Reflexionswege früher diffus und damit energiereicher auftreten als bei größeren Räumen (z. B. Treppenhaus TU Dresden), wo sie anfänglich nur diskret sind. Daher steigt C_{50} bzw. C_{80} gegenüber SRR bei $T_{60} = 700$ bzw. 1100 ms stärker als bei $T_{60} = 2000$ ms.

Weiterführend soll hier bereits auf die Ergebnisse der Experimente in Abschnitt 4.5

und 4.6 vorgegriffen werden, um die SRR- und C_{50}-Werte mit den zugehörigen erzielten Erkennungsraten ins Verhältnis zu setzen. Wie die Experimente in Abbildung 4.9 (a1) bzw. 5.3 (a) zeigen, wäre anzunehmen, dass die Störung der Sprache oder der Spracherkennung mit steigendem T_{60} ebenfalls steigt. C_{50} und C_{80} korrespondieren auch (im Wesentlichen) mit der Erkennungsrate aus den benannten Abbildungen, zumindest insofern, dass ein sinkendes C_{50}/C_{80} auch zu einer sinkenden Erkennungsrate führt. Beim SRR ist dies wie erwähnt nicht so. Dies stützt auch die nach den Experimenten aus Abbildung 4.11 (a) und (b) aufgestellte Theorie, dass sich Reflexionen mit einer Verzögerungszeit unter 50 ms nützlich auf die Spracherkennung auswirken. Im Vergleich der Erkennungsraten aus Abbildung 4.9 (a1) und den Maßen in Abbildung 4.8 oben scheint C_{50} (C_{80} schneidet sogar noch etwas besser ab) das geeignetere Störmaß im Vergleich zum SRR zu sein. Allerdings korrespondieren auch C_{50} und C_{80} nicht vollständig mit den Erkennungsraten. Außerdem wird der sehr starke Abfall der Erkennungsrate für den Wert $T_{60} = 2000$ ms in C_{50}/C_{80} in keiner Weise abgebildet.

Es wird daher an dieser Stelle vermutet, dass ein geeignetes Maß Reflexionen mit besonders kritischen Verzögerungszeiten stärker gewichten muss als andere. Diese können eventuell Bereiche zwischen (100 … 300) ms sein, was sich aus den Experimenten in Abbildung 4.11 (a) und (b) motiviert. Hier nicht vorgestellte Experimente zeigen allerdings, dass damit der sehr starke Effekt von T_{60} nach wie vor nicht genügend berücksichtigt wird, die Verläufe ähneln denen für C_{50}. In Abbildung 4.11 (b) zeigt sich im Fall $T_{60} = 2000$ ms, dass bereits geringfügige, aber sehr späte Reflexionen ($\approx$ (1500 … 2000) ms), die Erkennungsrate stark beeinflussen. Es wird, wie in Abschnitt 4.6.2 ausgeführt, die Vermutung angestellt, dass der störende Einfluss mit der Verzögerung der Reflexionen steigt. Sie sollten also in einem Störmaß mit steigender Verzögerungszeit mehr und mehr gewichtet werden. Frühe Reflexionen (< 50 ms) werden wie erläutert als nützlich angesehen. Um beide Effekte in ein geeignetes Störmaß einfließen zu lassen, kann bspw. ein

$$C_{\mathrm{ASR}} = 10\lg \frac{\int_0^{t_k=50\ \mathrm{ms}} h^2(t)\mathrm{d}t}{\int_{t_k=50\ \mathrm{ms}}^{\infty} (w_{C_{\mathrm{ASR}}} \cdot h)^2(t)\mathrm{d}t}\ \mathrm{dB} \tag{4.70}$$

gebildet werden, das die störenden Reflexionen mit der Wichtungsfunktion

$$w_{C_{\mathrm{ASR}}} = 100 \cdot (t - t_k)^2 \tag{4.71}$$

mit größer werdender Verzögerungszeit mehr und mehr anhebt. Die hier vorgestellte quadratische Wichtungsfunktion wurde aufgrund von einigen Experimenten einer linearen Wichtungsfunktion vorgezogen. Der Wert 100 ist ebenfalls experimentell ermittelt worden. Alle Annahmen sind Heuristiken. Gleichung (4.70) soll keine Lösung des Problems darstellen, sondern durch die Werte in Abbildung 4.8 unten andeuten, dass für ein C_{ASR} bspw. nach Gleichung (4.70) nicht einfach der SRR bzw. C_{50} bemüht werden kann. Es besteht daher die Notwendigkeit, ein C_{ASR} zu entwickeln. Dafür sind weitere Versuche nötig, das leistet diese Arbeit nicht.

4.5 Experimente zu Auswirkungen von Hall auf ASR

Im Vorfeld wurde der Einfluss des Halls auf die Sprache beschrieben. Ziel des folgenden Abschnittes ist es nun, die Auswirkungen von Hall auf ASR zu untersuchen. Zunächst wird beschrieben, wie unterschiedliche Raumumgebungen simuliert werden. Anschließend folgen Experimente zur Spracherkennung.

4.5.1 Messung der RR in simulierten Umgebungsbedingungen

In den folgenden Abschnitten werden Abhängigkeiten der Erkennungsrate von unterschiedlichen Raumbedingungen gemessen. Um entspechende Abhängigkeiten zu untersuchen, wird die Erkennungsrate für eine bestimmte Hallbedingung i gemessen und mit den Resultaten anderer Bedingungen verglichen. Für die Hallbedingung i erhält man den Messwert für die Erkennungsrate, indem jede Datei des Evaluationskorpus (hier Subset des APOLLO-Korpus, vgl. Abschnitt 2.3.6) mit der Bedingung i verhallt wird und ein Erkennungslauf über das Korpus durchgeführt wird. Das Verhallen geschieht durch Faltung der sauberen Sprachsignale $s(t)$ mit der RIR $h_i(k)$ der Bedingung i. Die benutzten RIRs stammen aus den in den Abschnitten 3.4.2 und 3.4.3 beschriebenen Messreihen. Im Wesentlichen teilen sich die Untersuchungen in die Abhängigkeit der Erkennungsrate von T_{60} und vom SMD auf. Letzteres stellt eine Neuheit dar. Üblicherweise wird in der Literatur nur T_{60} als Störparameter verändert.

4.5.2 Abhängigkeit der RR von T_{60}

Mit dieser Untersuchung soll die Abhängigkeit der Erkennungsrate RR von der Halligkeit von Räumen beschrieben werden. Die Halligkeit in Räumen hängt, wie in Abschnitt 3.2.2 dargestellt, von verschiedenen Faktoren und Größen ab. Diese sind insbesondere das Volumen V sowie die äquivalente Absorptionsfläche A. Allgemein lässt sich die Halligkeit eines Raumes mit der Nachhallzeit T_{60} beschreiben. Abhängigkeiten von V und A sind darin bereits enthalten. Nicht enthalten ist eine Abhängigkeit vom SMD. Dieser wäre streng genommen noch mit dem Hallradius ins Verhältnis zu setzen. Diese Betrachtung wird in dieser Arbeit jedoch nicht durchgeführt. Des Weiteren wird der Einfluss der in Abbildung 3.11 nachgewiesenen Frequenzabhängigkeit von T_{60} auf die RR in diesem Experiment nicht weiter beachtet. Es wird hier allerdings darauf hingewiesen, dass die für die folgenden Messungen benutzten RIRs in massiven Gebäuden aufgenommen wurden. Sie besitzen eine mit steigender Frequenz sinkende Nachhallzeit und ähneln qualitativ der RIR aus Abbildung 3.9 (a). Ihre Nachhallzeit (Vollband) variiert zwischen $0 \leq T_{60} \leq 2,0$ s. Die Erkennungsrate wird bei den folgenden konstanten SMDs ermittelt:

- **SMD = 100 cm** - Simulation von Nahfeldbedingungen. Bei diesem SMD wird angenommen, dass sich das Mikrofon noch im Hallradius befindet. Dies ist bei

entsprechender Richtwirkung durch den menschlichen Körper (vgl. Abbildung 3.2) normalerweise gegeben. 100 cm wurden außerdem gewählt, da es sich um eine typische Dialogentfernung handelt, d. h., ein Sprachbediener würde sich intuitiv nur bei Nichtfunktion oder starken Störgeräuschen näher zum Mikrofon neigen.

- **SMD = 200 cm** - Stellt zunächst einen weiteren Messpunkt zwischen Nah- und Fernfeldbedingung dar. Die Ergebnisse in Abbildung 4.9 lassen allerdings darauf schließen, dass der SMD von 200 cm bereits im Fernfeld liegt (Der Messaufbau bei den RIR-Messungen beinhaltete keine Maßnahmen zur Steigerung der Richtcharakteristik wie Abschirmungen etc. Es wird von einer Kugelcharakteristik ausgegangen.). Dieser Abstand wird im späteren Teil der Experimente nicht weiter berücksichtigt.
- **SMD = 300 cm** - Simulation von Fernfeldbedingungen. Bei diesem Abstand wird angenommen, dass sich das Mikrofon für die meisten hier betrachteten Räume außerhalb des Hallradius befindet (vgl. Abbildung 3.12 (c) bzw. Abbildung 3.2). Ein SMD = 300 cm ist außerdem nahe der Obergrenze für Sprachbedienungen in Wohnräumen (vgl. Abschnitt 3.4.2 bzw. Tabelle 3.3).

Die Ergebnisse der Messungen sind in Abbildung 4.9 (a1) dargestellt. Die Erkennungsrate nimmt mit steigendem T_{60} rapide ab. Zusätzlich wird die Abhängigkeit vom SMD deutlich, wobei RR für 200 und 300 cm dicht bei einander liegen. Daraus kann geschlussfolgert werden, dass SMD = 200 cm der Fernfeldbedingung bereits sehr nahe kommt. Allgemein lässt sich feststellen, dass bereits in gering halligen Umgebungen, wie dem Wohn- und Büroumfeld ($0,3\text{ s} < T_{60} < 0,8\text{ s}$, vgl. Abschnitt 3.4.2), die Erkennungsleistung rapide abnimmt. Bei Umgebungen wie Treppenhäusern (Bsp. hier $T_{60} = 2,0\text{ s}$) ist mit dieser Erkennerkonfiguration keine sinnvolle Erkennung mehr möglich.

4.5.3 Abhängigkeit der RR vom SMD

Mit dieser Untersuchung soll die Abhängigkeit der Erkennungsrate von der Sprecherposition beschrieben werden, die durch den SMD simuliert werden kann. Die Abhängigkeiten vom SMD werden anhand der im SMART-Room gemessenen RIRs ermittelt. Bei diesen Experimenten wird die Nachhallzeit konstant belassen, was aufgrund des gleichen Raumes gegeben ist. T_{60} beträgt ca. 650 ms. Der Raum eignet sich daher gut als (eher halliger) Repräsentant von Räumen im Wohnumfeld ($0,3\text{ s} < T_{60} < 0,8\text{ s}$, vgl. Abschnitt 3.4.2).

Am Verlauf der Messergebnisse in Abbildung 4.9 (a2) kann das Verhalten der direkten und diffusen Energieanteile beobachtet werden. Wie in Abbildung 3.2 angedeutet, sinkt der DRR bereits nach wenigen cm SMD, sodass sich die Erkennungsrate rapide verschlechtert. Nach ca. (100 ... 200) cm kommt der SMD in die Nähe des Hallradius

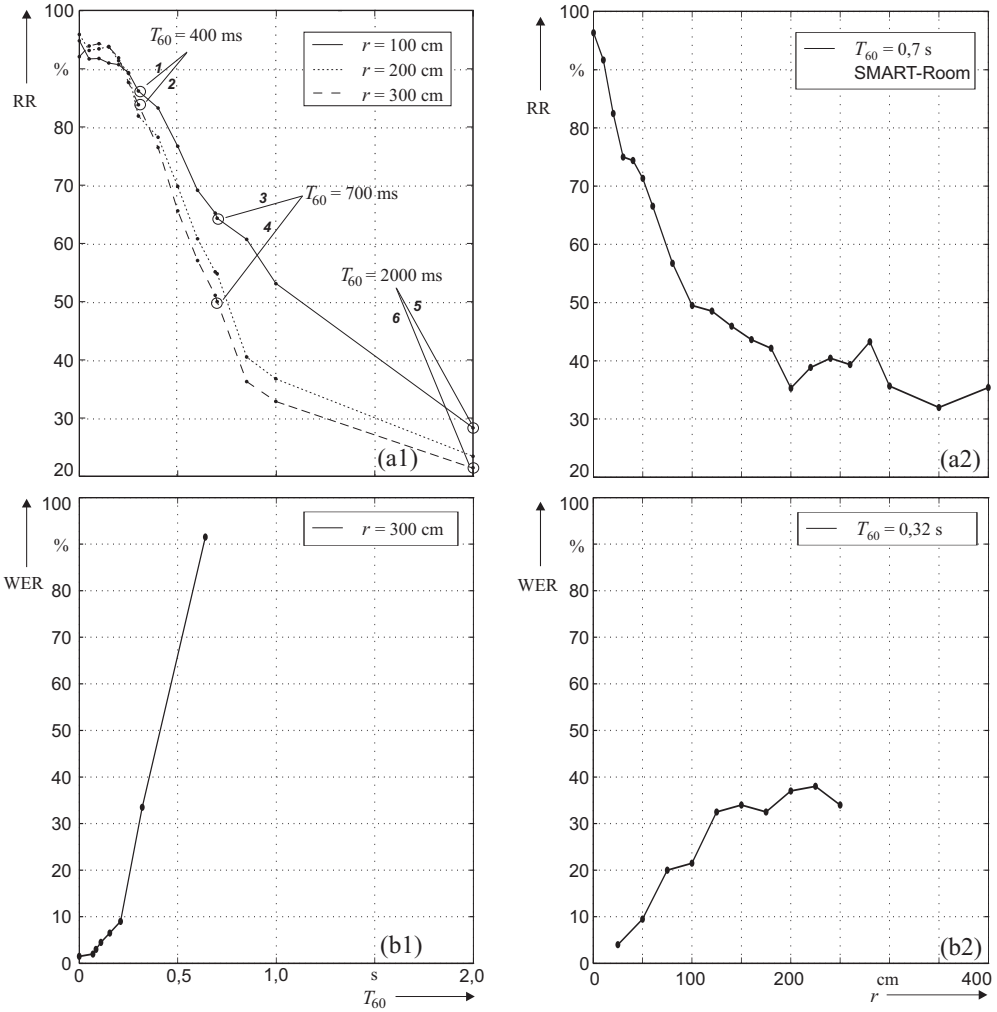

Abbildung 4.9 – Messung der Erkennungsrate in Abhängigkeit von T_{60} und SMD.
Oben: Messungen in dieser Arbeit, Subset des APOLLO-Kopus und UASR-Spracherkenner.
(a1) RR abhängig von T_{60} beim SMD = (100; 200; 300) cm. Die markierten Arbeitspunkte beziehen sich auf die Experimente in Abschnitt 4.6.
(a2) RR abhängig vom SMD bei $T_{60} = 0,7$ s, SMART-Room-Umgebung [NGMH07].
Unten: Zum Vergleich Messungen von Emanuel Habets. Die Graphiken wurden aus [Hab06] übernommen, wurden aber für diese Vergleichsdarstellung so angepasst, dass die Werte der Abszissen mit der oberen Graphik übereinstimmen.
(b1) RR abhängig von T_{60} bei einem SMD = 300 cm.
(b2) RR abhängig vom SMD bei $T_{60} = 0,32$ s.

und die Erkennungsrate nähert sich dem Wert für das Fernfeld an (Im Gegensatz zu den anderen RIR-Messungen wurde im SMART-Room mit menschlicher Körperabschattung gemessen, wodurch Richtcharakteristik und Hallradius erhöht sind.). Die Messungen zeigen eine deutliche Abhängigkeit der Erkennungsrate von der Sprecherposition und machen deutlich, dass die Nachhallzeit als alleiniges Kriterium für die Halligkeit eines bestimmten Szenarios nicht ausreichend ist.

4.5.4 Vergleichende Bewertung der Ergebnisse

Um die Sinnfälligkeit der Ergebnisse sicherzustellen, werden sie zusätzlich mit Messergebnissen von Emanuel Habets [Hab06] verglichen (Abbildung 4.9 (b1) und (b2)), der damit seine Forschungen zur blinden Enthallung motiviert. Habets misst anstelle der RR die WER. Seine Experimente basieren nicht auf einer Kommandoworterkennung, sondern auf einem Large-Vocabulary-Continious-ASR-System. Das erklärt, warum die WER sensibler reagiert als die hier gemessene RR. Weitere Autoren berichten von ähnlichen Abhängigkeiten, allerdings wird oft nur die Nachhallzeit als veränderliche Größe angegeben und die Angabe eines SMD vollständig ausgelassen.

4.6 Einfluss verschiedener Hallkomponenten auf die Spracherkennung

In den zuvor beschriebenen Experimenten wird der störende Einfluss des Raumhalls auf die Spracherkennung dargestellt. Ziel dieses Abschnittes ist es zu untersuchen, welche Hallkomponenten besonders störend wirken bzw. ob, ähnlich wie bei der menschlichen Sprachverständlichkeit, nützliche Hallphasen existieren (vgl. Abschnitt 4.4). Diese Experimente dienen der Vorbereitung der Entwicklung von geeigneten Ansätzen, die Robustheit gegen die Hallstörung erreichen sollen. Im Gegensatz zu Ansätzen, die versuchen, den Hall vollständig zu eliminieren, besteht hier die Erwartung, dass es ausreicht, nur bestimmte, störende Teile zu eliminieren. Der Hintergedanke ist dabei, dass sich die Komplexität bereits bestehender Ansätze (vgl. Abschnitt 5) verringert, wenn nur ein Teil der Hallabschnitte eliminiert werden muss, sodass zu hohe Rechenleistungsanforderungen bzw. Adaptionszeiten der aktuellen Ansätze in einen benutzbaren Bereich gelangen.

4.6.1 Modifikation von RIRs

Für die Experimente werden die RIRs modifiziert, indem die Hallphasen systematisch beschnitten werden, um den Einfluss bestimmter Hallabschnitte zu ermitteln. Als Ausgangspunkt dienen dabei RIRs sechs ausgewählter Bedingungen aus dem Experiment

Tabelle 4.3 – Ausgangspunkt: originale RIRs für die folgenden Experimente mit modifizierten RIRs. Die RIRs korrespondieren mit den markierten Bedingungen aus Abbildung 4.9 (a1).

	SMD	T_{60}	Raumtyp
1	100 cm	400 ms	Wohnzimmer
2	300 cm	400 ms	Wohnzimmer
3	100 cm	700 ms	Büroraum
4	300 cm	700 ms	Büroraum
5	100 cm	2.000 ms	Treppenhaus
6	300 cm	2.000 ms	Treppenhaus

in Abbildung 4.9 (a1) (durch Kreis markiert). Die ausgewählten Bedingungen sind in Tabelle 4.3 aufgelistet. Die zugehörigen RIRs werden so beschnitten, dass nur bestimmte Hallphasen in der RIR verbleiben. Dazu werden in jeder RIR $h(t)$ zunächst die Direktschallphase $h_{\mathrm{D}}(t)$, repräsentiert durch den Impuls bei t_0, und die Hallphase $h_{\mathrm{R}}(t) = h(t > t_0)$ voneinander getrennt. Die Extraktion erfolgt unter Benutzung eines Von-Hann-Fensters

$$w_{\mathrm{Hann}}(t) = \begin{cases} 0 & ; \quad t < -T_w/2 \\ \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{T_w}t\right) & ; \quad -T_w/2 \leq t \leq +T_w/2 \\ 0 & ; \quad t > +T_w/2 \end{cases} \tag{4.72}$$

mit

$$h_{\mathrm{D}}(t) = \begin{cases} 0 & ; \quad t < t_0 - T_w/2 \\ h(t) \cdot w_{\mathrm{Hann}}(t - t_0) & ; \quad t_0 - T_w/2 \leq t \leq t_0 + T_w/2 \\ 0 & ; \quad t > t_0 + T_w/2 \end{cases}, \tag{4.73}$$

$$h_{\mathrm{R}}(t) = \begin{cases} 0 & ; \quad t < t_0 \\ h(t) \cdot w_{\mathrm{Hann}}(t - t_0 - T_w/2) & ; \quad t_0 \leq t \leq t_0 + T_w/2 \\ h(t) & ; \quad t > t_0 + T_w/2 \end{cases}, \tag{4.74}$$

wobei T_w die Breite des Fensters darstellt; in den Experimenten gilt $T_w = 32$ ms. Die Modifikationen der RIR werden nur an $h_{\mathrm{R}}(t)$ vorgenomen, $h_{\mathrm{D}}(t)$ hingegen repräsentiert das direkte Signal und bleibt unverändert. Abbildung 4.10 beschreibt schematisch die vier angewandten Modifikationen von h_{R}:

(a) Entfernung später Reflexionen: Das Ende von $h_{\mathrm{R}}(t)$ wird nach einer Zeit $t_{\mathrm{cut\,off}}$ auf 0 gesetzt (vgl. Abbildung 4.10 (a)). Damit am Schnitt keine Sprünge und somit ungewollte Frequenzanteile entstehen, wird der Schnitt durch ein Ausblendefenster geglättet. Zum Ausblenden wird $h_{\mathrm{R}}(t)$ an der Schnittstelle mit der zweiten

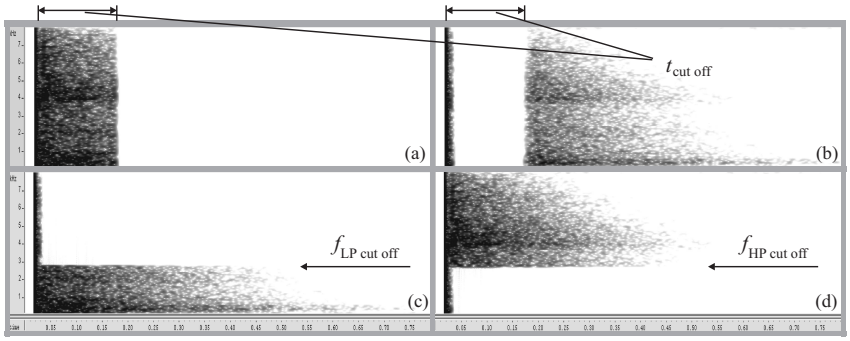

Abbildung 4.10 – Qualitative Graphik zur Darstellung der vier vorgenommen Modifikationen der RIRs:
(a) Entfernung später Reflexionen nach der Zeit $t_{\text{cut off}}$
(b) Entfernung früher Reflexionen vor der Zeit $t_{\text{cut off}}$
(c) Entfernung hochfrequenter Reflexionen mit Frequenzen über $f_{\text{LP cut off}}$
(d) Entfernung tieffrequenter Reflexionen mit Frequenzen unter $f_{\text{HP cut off}}$.

Hälfte von $w_{\text{Hann}}(t - t_{\text{cut off}} + T_w/2)$ multipliziert. $t_{\text{cut off}}$ beschreibt demnach die Zeit zwischen t_0 und dem Ende des Ausblendens.

(b) Entfernung früher Reflexionen: Der Anfang von $h_{\text{R}}(t)$ wird bis zu einer Zeit $t_{\text{cut off}}$ auf 0 gesetzt (vgl. Abbildung 4.10 (b)). Damit am Schnitt keine Sprünge und somit ungewollte Frequenzanteile entstehen, wird der Schnitt durch ein Einblendefenster geglättet. Zum Einblenden wird $h_{\text{R}}(t)$ an der Schnittstelle mit der ersten Hälfte von $w_{\text{Hann}}(t - t_{\text{cut off}} - T_w/2)$ multipliziert. $t_{\text{cut off}}$ beschreibt demnach die Zeit zwischen t_0 und dem Beginn des Einblendens.

(c) Entfernung hochfrequenter Reflexionen: $h_{\text{R}}(t)$ wird unter Benutzung eines FIR-Filters tiefpassgefiltert. Die Stoppbanddämpfung beträgt 60 dB. Das Stoppband beginnt 50 Hz oberhalb einer Frequenz $f_{\text{LP cut off}}$ (vgl. Abbildung 4.10 (c)).

(d) Entfernung tieffrequenter Reflexionen: $h_{\text{R}}(t)$ wird unter Benutzung eines FIR-Filters hochpassgefiltert. Die Stoppbanddämpfung beträgt 60 dB. Das Stoppband beginnt 50 Hz unterhalb einer Frequenz $f_{\text{HP cut off}}$ (vgl. Abbildung 4.10 (d)).

Um die modifizierte RIR $h_{\text{mod}}(t)$ zu erhalten, wird die modifizierte Hallphase $h_{\text{R,mod}}(t)$ in allen vier Fällen zu $h_{\text{D}}(t)$ addiert, sodass

$$h_{\text{mod}}(t) = h_{\text{D}}(t) + h_{\text{R,mod}}(t), \tag{4.75}$$

wie in den Abbildungen 4.10 (a) – (d) dargestellt, entsteht. D. h., nach der Faltung

$$x_{h_{\text{mod}}}(t) = (h_{\text{mod}} * s)(t) \tag{4.76}$$

ergibt sich mit $x_{h_{\text{mod}}}(t)$ ein Signal, das die vollständige Direktschallkomponente und die in der RIR belassene Hallkomponente enthält.

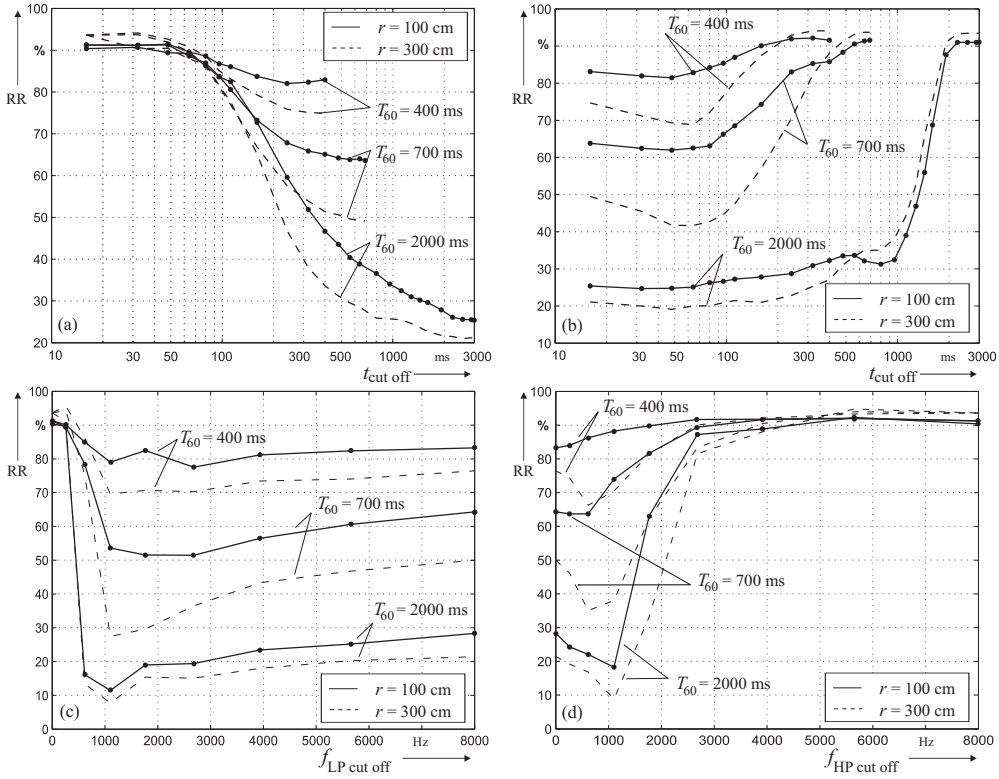

Abbildung 4.11 – Einfluss von (a) frühen, (b) späten, (c) tieffrequenten und (d) hochfrequenten Reflexionen auf die Erkennungsleistung. Die Beschneidung der Hallphase richtet sich nach Abschnitt 4.6.1.

4.6.2 Einfluss früher und später Reflexionen auf die RR

Um den Einfluss früher und später Reflexionen auf die Erkennungsleistung herauszufinden, wird bei den sechs ausgewählten RIRs nach Tabelle 4.3 die Hallphase zeitlich modifiziert.

Abbildung 4.11 (a) zeigt mit dem Experiment zur Eliminierung später Reflexionen nach Abbildung 4.10 (a), dass die Erhöhung von $t_{\text{cut off}}$ erwartungsgemäß zu einer Verringerung der Erkennnungsrate führt. Zwischen etwa 80 ms $< t_{\text{cut off}} <$ 300 ms ist die größte Änderung zu beobachten, woraus geschlussfolgert werden kann, dass Reflexionen mit Verzögerungen in diesem Bereich besonders störend wirken. Reflexionen mit geringeren Verzögerungszeiten als 50 ms können nicht als störend nachgewiesen werden. Reflexionen zwischen 50 und 80 ms wirken sich geringfügig störend aus.

Abbildung 4.11 (b) zeigt mit dem Experiment zur Eliminierung früher Reflexionen nach Abbildung 4.10 (b), dass, beginnend mit $t_{\text{cut off}} = \infty$, die Verringerung von $t_{\text{cut off}}$ ebenfalls erwartungsgemäß zu einer Verringerung der Erkennnungsrate führt.

In beiden Experimenten bildet die Bedingung Treppenhaus ($T_{60} = 2000$ ms) eine Ausnahme; sie stellt eine Extrembedingung dar. In Abbildung 4.11 (b) zeigt sich im Fall $T_{60} = 2000$ ms, dass bereits geringfügige, aber sehr späte Reflexionen ($\approx$ (1500 ... 2000) ms), die Erkennungsrate stark beeinflussen. Es wird demnach die Vermutung angestellt, dass der störende Einfluss mit der Verzögerung der Reflexionen steigt, da diese späten Reflexionen nur sehr geringe Energie besitzen[16]. Für die beiden anderen Räume, die mit ihren Nachhallzeiten gerade den Bereich des Wohn- und Büroumfeldes abstecken (0,3 s $\leq T_{60} \leq$ 0,8 s), wird festgehalten, dass Reflexionen mit Verzögerungszeiten im Bereich 80 ms $< t_{\text{cut off}} <$ 300 ms als besonders störend angesehen werden.

In beiden Darstellungen ist zusätzlich zu beobachten, dass die Erkennungsrate ohne Hallkomponente für den SMD = 300 cm geringfügig besser ist als für den SMD = 100 cm (in Abbildung 4.11 (a) ganz links bzw. in Abbildung 4.11 (b) ganz rechts). Bei Betrachtung der beiden Graphiken entsteht dabei der Eindruck, dass frühe Reflexionen die Spracherkennung verbessern. Bei der Gewinnung von $h_{\text{D}}(t)$ nach (4.73) werden neben dem Impuls auch besonders frühe Reflexionen ($<$ 16 ms) extrahiert. Es besteht daher die Vermutung, dass diese den direkten Schall mit Energie anreichern und sich somit, ähnlich wie bei der menschlichen Wahrnehmung, scheinbar auch für die Spracherkennung als nützlich erweisen. Dieser verstärkende Effekt ist bei der Fernfeldbedingung SMD = 300 cm größer, da der Direktschall schwächer als bei der Nahfeldbedingung SMD = 100 cm ist. Diese Theorie wird durch die Ergebnisse in Abbildung 4.11 (b) bei geringen Werten für $t_{\text{cut off}} < 60$ ms untermauert, wo die Erkennungsrate bei Hinzuziehen der frühen Reflexionen gesteigert wird. Besonders signifikant tritt dieser Effekt beim SMD = 300 cm auf, der damit konkret den frühen Reflexionen zugeordnet werden kann. Um genauere Schlussfolgerungen zu ziehen sind weitere Experimente nötig, die jedoch in dieser Arbeit nicht weiter ausgebaut werden.

4.6.3 Einfluss hoch- und tieffrequenter Reflexionen auf die RR

Um herauszufinden, ob es besonders störende Hallfrequenzen gibt, werden die Hallphasen der gleichen sechs Bedingungen der vorhergehenden Experimente gefiltert, sodass spezielle Frequenzen eliminiert werden. Dabei werden die Werte für $f_{\text{HP cut off}}$ bzw. $f_{\text{LP cut off}}$ in gleichen Schritten auf der Mel-Skala nach Gleichung (2.9) erhöht, um jeweils eine vergleichbare Änderung im Merkmalvektor (MFB-Ausgang) zu erreichen.

[16] Zur Erläuterung: Bei einer RIR mit $T_{60} = 2000$ ms ist die Energie bei $t = 2000$ ms theoretisch um 60 dB im Vergleich zur Energie bei Beginn der Hallphase (t nahe 0) gefallen. Bei 1500 ms beträgt der Unterschied immer noch 45 dB. Beide Werte stellen in ihrer Größenordnung für die Spracherkennung normalerweise keine Störung dar. SNR = 25 dB gilt als nahezu ungestört.

Abbildung 4.11 (c) zeigt das Experiment zur Eliminierung hochfrequenter Reflexionen nach Abbildung 4.10 (c). Die Ergebnisse beschreiben einen sehr starken Abfall bei bereits geringer Erhöhung von $f_{\text{LP cut off}}$ (oberhalb von 250 Hz). Bei weiterer Erhöhung von $f_{\text{LP cut off}}$ oberhalb von 2500 Hz zeigt sich eine leichte Verbesserung der Erkennungsrate. Dieser Effekt kann damit erklärt werden, dass insbesondere Frikative und ähnliche Laute gestärkt werden und somit eine Steigerung von RR bewirkt wird. Sie bestehen in oberen Frequenzen vorwiegend aus verschiedenartigem Rauschen, welches durch das Hinzufügen von hochfrequenten Reflexionen mit Energie angereichert wird (vgl. Beschreibung in Abschnitt 4.3.1). Dieser Effekt wird in dieser Arbeit nicht genauer untersucht.

Abbildung 4.11 (d) zeigt das Experiment zur Eliminierung tieffrequenter Reflexionen nach Abbildung 4.10 (d). Betrachtet man die Ergebnisse beginnend mit $f_{\text{HP cut off}} = 8000$ Hz, so bleibt die Verringerung von $f_{\text{HP cut off}}$ bis auf 2500 Hz (Hinzufügen hochfrequenter Reflexionen) ohne wesentliche Einflüsse auf die Erkennungsrate; sogar bei der Extrembedingung Treppenhaus. Man kann demnach davon ausgehen, dass hochfrequenter Hall (> 2500 Hz) wenig bis gar keinen störenden Einfluss besitzt. Das ist die für diese Arbeit wichtigste Erkenntnis der Experimente mit modifizierten RIRs. Dieser Effekt kann auf die folgenden Ursachen zurückgeführt werden:

- Tiefe Frequenzen hallen in massiven Gebäuden stärker als hohe (vgl. Abbildung 3.9 (a) bzw. 3.11).
- Das Verzerren von Frikativrauschen durch Hall führt wieder zu Rauschen, und demnach kaum zu einem Unterschied im Frikativ selbst.
- Die zeitliche Verschleifung eines Frikativs in den folgenden stimmhaften Laut hinein besitzt im Vergleich zu diesem nur eine geringe Energie, da die Energie von stimmhaften Lauten im Normalfall wesentlich größer ist als die von stimmlosen (vgl. Sonoritätsklassen in Tabelle 4.2 und Beschreibung in Abschnitt 4.3.1). Somit ist die Störung nur gering.

Im Unterschied dazu entsteht bei weiterer Verringerung von $f_{\text{HP cut off}} < 2500$ Hz ein starker Abfall der Erkennungsrate. Tieffrequenter Hall wirkt folglich besonders störend auf die Spracherkennung. Auch dieser Effekt ist mit der Sonorität (Tabelle 4.2) zu erklären: Aufgrund der wesentlich größeren Energie der stimmhaften Laute (im Vergleich zu stimmlosen), die besonders in den tieffrequenteren Regionen im Spektrum verteilt ist, ist deren Hall ebenfalls besonders energiereich. Interessant ist, dass die Störung besonders bei den Formantfrequenzen der ersten drei Formanten zu finden ist. Die in Abschnitt 4.3.1 beschriebene Theorie, dass die in einen sich anschließenden Frikativ hineinragende Hallfahne eines stimmhaften Lautes (besonders der energiereichen Formanten) den energetisch schwächeren Frikativ erheblich stark stört, wird durch diese Ergebnisse unterstützt. Im Übrigen wird auch ein sich anschließender stimmhafter Laut bereits signifikant gestört.

4.7 Zusammenfassung der Erkenntnisse

Nach Darstellung von relevanten Aspekten der menschlichen Sprache beschreibt dieses Kapitel, wie sich der Einfluss des Raumes auf die Sprache auswirkt. Dabei wird ein Schwerpunkt auf die Veränderung der zeitlichen Modulation gelegt. Im Anschluss werden wichtige Aspekte des klassischen Forschungsgebietes des Einflusses des Raumes auf die menschliche Sprachverständlichkeit gegeben. Es werden typische Maße vorgestellt, woraus versucht wird, ein Störungsmaß für die Spracherkennung abzuleiten. Nach diesen Vorbetrachtungen werden Experimente vorgestellt, die den störenden Einfluss des Raumes auf die Spracherkennung herausstellen. Die Experimente unterteilen sich in Abhängigkeiten von der Nachhallzeit und vom SMD. Um herauszufinden, welche Bereiche des Halls besonders störend wirken, werden weitere Experimente durchgeführt, die mit modifizierten RIRs arbeiten. Bei der Modifikation wird die Hallphase der RIR jeweils so beschnitten, dass nur bestimmte Hallbereiche in der RIR verbleiben. Aus den Erkennungsraten können interessante Schlussfolgerungen gezogen werden. Folgende Erkenntnisse werden gewonnen:

- Modulationsfrequenzen, die geringer als 20 Hz sind, sind für die Sprachverständlichkeit sowie für die Spracherkennung wichtig.
- Modulationsfrequenzen, die größer als 20 Hz sind, können als Störung angesehen und demnach für die Spracherkennung eliminiert werden.
- SRR sowie C_{50} sind nicht geeignet, um die Störung für die Spracherkennung zu beschreiben.
- Frühe Reflexionen (< 50 ms) wirken sich nicht störend auf die Spracherkennung aus. Sie haben scheinbar einen positiven Effekt.
- Reflexionen mit einer Verzögerung unter 50 ms wirken sich nicht störend auf die Spracherkennung aus.
- Reflexionen mit einer Verzögerung zwischen (50 ... 80) ms wirken sich gering störend auf die Spracherkennung aus.
- Reflexionen mit einer Verzögerung über 80 ms wirken sich störend auf die Spracherkennung aus.
- Bei welchen Verzögerungen sich Reflexionen am stärksten störend auswirken kann diese Arbeit nicht beantworten.
- Hochfrequenter Hall (> 2500 Hz) ist nahezu harmlos für die Spracherkennung.
- Tieffrequenter Hall (< 2500 Hz) wirkt besonders störend auf die Spracherkennung.
- Sehr tieffrequenter Hall (< 250 Hz) wirkt nicht störend auf die Spracherkennung.

5 Existierende Lösungsansätze

5.1 Überblick

Dieses Kapitel gibt einen Überblick über Lösungsansätze, die zur Robustheit von Spracherkennungssystemen unter Hallbedingungen beitragen können. Die Unterteilung der Ansätze wird dabei an die Einteilung der Maßnahmen gegen additives Rauschen angelehnt. Es entstehen die vier Ansatzebenen akustische Ebene, Signalebene, Merkmalebene und Modellebene, auf denen Maßnahmen gegen Störungen integriert werden können.

5.2 Robustheitssteigernde Maßnahmen bei ASR-Systemen

Der wichtigste Grund, warum sich die Erkennungsleistung bei ASR-Systemen unter realen akustischen Umgebungsbedingungen verringert, ist:

das Abweichen der Merkmalmuster des Testsignals von denen des Trainingssignals.

Die Abweichungen werden vorwiegend durch Störungen im Signal eingebracht, die im Trainingsmaterial nicht vorhanden sind. Dies kann z. B. in Abbildung 6.4 beobachtet werden. Zusätzlich existieren sprecherbezogene Abweichungen. Folgende Störeinflüsse können festgestellt werden:

- **Variabilität der Sprache** – klassisches Problemfeld in der Spracherkennung, das durch die gegebenen Anforderungen an Spracherkenner bestimmt ist und bereits enorme Schwierigkeiten hervorruft. Es wird in dieser Arbeit nicht untersucht. Variabilität der Sprache entsteht durch:
  - Variabilität bei einem Sprecher (Sprechtempo, Emotion, Lombardeffekt usw.),
  - Variabilität bei mehreren Sprechern (Dialekt, Akzent, Alter, Geschlecht usw.),
  - kontextbezogene Variabilität (Diktat, Dialog, Konversation usw.).
- **additive Störungen** – entstehen durch fremde Schallquellen, deren Signale sich am Mikrofon additiv einkoppeln. Sie können unterteilt werden in:

  - stationäre Geräusche (Lüfterrauschen usw.),
  - nichtstationäre Geräusche (konkurrierende Sprecher, Radioton, Türenklappen usw.) und
  - systemeigene Geräusche (Es sind von einem Lautsprecher des Gerätes ausgegebene Töne. Sie gehören normalerweise zu den nichtstationären Geräuschen, erhalten aber eine Sonderrolle, da das Quellsignal im System bekannt ist. Es entstehen sogenannte Rückkopplungen oder auch akustische Echos. Gegen diese Art von Störungen existiert eine besondere Gruppe von Methoden, die akustische Echokompensation (AEC)).

- **convolutive Störungen** – entstehen durch die Faltung der Impulsantwort des Übertragungssystems zwischen Sprecher und Mikrofon mit dem originalen Sprachsignal. Sie sind also vom Sprachsignal abhängig und damit instationär. Sie können unterteilt werden in:
  - Raumhall (lange Impulsantworten, hallend) und
  - Kanalverzerrungen (kurze Impulsantworten, färbend, im einfachsten Beispiel durch ein nicht ideales Mikrofon).

Die Steigerung der Robustheit von Spracherkennungssystemen gegen additive Störungen ist, wie in Abschnitt 1.1 erwähnt, bereits ein traditionelles Forschungsgebiet. Diese Maßnahmen setzen auf verschiedenen Ebenen im Erkennungssystem an, um die Abweichung der Testmuster von den Trainingsmustern zu verringern [Gon95]. Das sind:

- akustische Ebene,
- Signalebene,
- Merkmalebene und
- Modellebene.

Betrachtet man Abbildung 2.1, so findet man die Ebenen in der hierarchischen Struktur eines ASR-Systems wieder. Es existiert eine große Anzahl von Ansätzen gegen additives Rauschen, die in vielen Arbeiten überblicksartig dargestellt werden. Als Standardwerk galt lange das Buch von Junqua [JH95] (1995), das aufgrund der rasanten Entwicklungen auf diesem Gebiet derzeit nicht mehr auf dem neusten Stand ist. In der Zwischenzeit entstanden viele weitere Arbeiten, die den Stand der Technik für Maßnahmen gegen Rauschen darstellen (z. B. [Jos02]). Ein aktueller Überblick wird in [Dro08] gegeben.

Aufgrund der Unterschiedlichkeit von additiven und convolutiven Störungen ist der Einsatz der Maßnahmen, die gegen additives Rauschen entwickelt wurden, bei Raumhall meist nicht möglich. Trotzdem soll die typische Einteilung dieser Maßnahmen als Vorbild für die Strukturierung von existierenden Maßnahmen gegen Raumhall, um die es in dieser Arbeit geht, gelten. Ansätze gegen Raumhall reichen nicht wesentlich weiter als 10 Jahre zurück. Zunächst entwickeln sich Ansätze der blinden Enthallung von Sprachsignalen. Dazu existieren bereits einige wenige Arbeiten, die einen Überblick zum Stand der Technik geben [NG05, Hab06, HBC08, NJY+08]. Blinde Enthallung

ist ein eigenes Forschungsgebiet und hat nicht notwendigerweise die Spracherkennung zur Anwendung, kann aber zu deren Robustheit beitragen. Zu einem eigenständigen Forschungsthema – der Robustheit von ASR-Systemen gegen raumakustische Umgebungsbedingungen – existieren nur vereinzelte Arbeiten. In [SK08] gibt Sehr u. a. einen umfangreichen Überblick zu robustheitssteigernden Ansätzen für ASR-Systeme in halligen Umgebungsbedingungen. Dabei unterteilt auch er die Ansätze in die oben benannten Ebenen, in die auch die Maßnahmen gegen Rauschen eingeteilt werden. Weitere Artikel, die einen Überblick zum Stand der Technik dieses Forschungsthemas präsentieren, sind dem Autor nicht bekannt. Auch in [BSH08a], wo der aktuelle Stand der Technik (2008) im übergeordneten Forschungsgebiet der digitalen Sprachsignalverarbeitung abgebildet ist, gibt es zu diesem Thema keinen Artikel, wohingegen ein eigenes Kapitel für die Robustheit gegen Umgebungsgeräusche [Dro08] erstellt worden ist. Das Thema besitzt deshalb aktuelle Brisanz. Die existierenden Lösungsansätze, die zur Steigerung der ASR-Robustheit gegen Hall möglich sind, werden im Folgenden kurz vorgestellt.

5.3 Akustische Ebene

Auf der akustischen Ebene wird versucht, Störungen so zu beeinflussen, dass sie im aufgenommenen Mikrofonsignal gar nicht erst enthalten sind. Diese Maßnahmen können u. U. sehr effektiv sein, da diese Störungen später durch die Signalverarbeitungsmethoden nicht mehr behandelt werden müssen. Eine grobe Einteilung kann in konstruktive Maßnahmen, die die Sprecher-Mikrofon-Strecke optimieren, sowie in Maßnahmen, die die Richtwirkung der Schallaufnahme des Mikrofons beeinflussen, erfolgen.

5.3.1 Konstruktive Maßnahmen

Im einfachsten Fall sind mechanische Dämpfungen, Abschirmungen usw. an einer Geräuschquelle vorzunehmen. Bei Raumhall existiert keine explizite Störquelle. Allerdings können die Reflexionspunkte (Gegenstände, Wände etc.) abstrakt als Störquelle gesehen werden. Eine Verbesserung der Absorptionsgrade der Reflexionspunkte (Teppichboden, Polstermöbel etc.) führt zur Verringerung der Nachhallzeit des Raumes und damit zu einer besseren Leistungsfähigkeit des Erkenners (vgl. Abbildung 4.9). Wird der Raum als gegeben und nicht veränderbar hingenommen, sind konstruktive Maßnahmen am Empfänger denkbar. Dazu zählt in erster Linie die Körperschallentkopplung des Mikrofons (gegen mechanische Vibrationen). Weiterhin schaffen Abschirmungen etc. die Möglichkeit, Schall aus der Richtung des Sprechers aufzunehmen und aus anderen Richtungen zu unterdrücken. Eine weitere konstruktive Maßnahme ist die Optimierung des SMDs, die z. B. durch die Benutzung von Headsets o. ä. erreicht werden kann. Dies ist allerdings, wie Abschnitt 1.1 erwähnt, nicht immer akzepta-

bel. Die Optimierung des SMDs lässt sich auch dadurch erreichen, dass an mehreren Punkten des Raumes (Tisch, Wände, Lampe o. ä.) Mikrofone angebracht werden, von denen bei verschiedenen Sprecherpositionen jeweils das mit dem aktuell kleinsten SMD zugeschaltet wird.

5.3.2 Richtwirkung der Schallaufnahme durch Richtmikrofone

Hall hat die Eigenschaft, eine räumlich diffuse Störung (idealisierte Annahme) zu sein, d. h., Schallwellen des Halls treffen aus allen Richtungen am Mikrofon ein. Gleiches gilt für diffuses Rauschen. Hat man die Möglichkeit, den Schall nur in Richtung des Sprechers aufzunehmen, reduziert sich die Hallstörung um die Dämpfung, die für die anderen Richtungen erreicht wird. Durch die konstruktive Maßnahme Abschirmung wurde bereits eine Richtwirkung der Schallaufnahme erzielt. Eine weitere Maßnahme ist die Benutzung von Richtmikrofonen. Aufgrund der geringen Kosten sind dabei Elektretkapseln mit Nierencharakteristik besonders wichtig. Eine bessere Richtwirkung kann mit Studiomikrofonen erzielt werden, die aber deutlich preisintensiver sind. Einen Überblick zu Mikrofonen findet man bspw. in [BP99, Dic99].

5.3.3 Räumliche Verarbeitung – Beamforming mit Mikrofon-Arrays

Es ist eine Frage der Betrachtungsweise, ob Beamformer als Methoden auf Signalebene angesehen werden können oder nicht. In dieser Arbeit sollen sie eine eigene Gruppe bilden, da die Methoden auf Signalebene versuchen, ein Signal zu enthallen. Ein Beamformer hingegen kann abstrakt als ein Mikrofon mit Richtwirkung gesehen werden, das Hall (bzw. Rauschen) von Seitenrichtungen gar nicht erst (bzw. nur gedämpft) aufnimmt. Somit kann ein Beamformer zur Steigerung der Robustheit unter Hallbedingungen beitragen.

Ansätze zu Beamformern können grob in fixe und adaptive Beamformer eingeteilt werden. Sie besitzen zunächst den Nachteil, dass sie mehrere Mikrofone benötigen, was zusätzliche Installations- und Hardwarekosten verursacht. Die Wissenschaft zu Beamforming befasst sich bislang vorwiegend mit der Robustheit gegen additives Rauschen. Robustheit gegen Hall ist gewöhnlich nicht der Fokus, wodurch ein Vergleich der Funktionsweise unter Hallbedingungen erschwert wird.

Fixe Beamformer Fixe Beamformer sind der klassische Delay-and-Sum-Beamformer (DSB) sowie der Filter-and-Sum-Beamformer (FSB) [BW01]. Sie werden durch mehrkanalige Mikrofonanordnungen (sogenannte Mikrofon-Arrays) gebildet, deren abgetastete Eingangssignale $x_m(k)$ so miteinander verrechnet werden, dass eine Richtcharak-

teristik entsteht. Das Ausgangssignal x_{DSB} des DSBs wird durch die Summe

$$x_{\mathrm{DSB}}(k) = \sum_{m=1}^{M} x_m(k - \kappa_\tau) \tag{5.1}$$

beschrieben, wobei $x_m(k - \kappa_\tau)$ das vom m-ten der M Mikrofone aufgenommene Signal darstellt, das um die Zeit τ (diskrete Zeit κ_τ) verzögert (engl.: delayed) wird. Durch die Einstellung von κ_τ kann die Richtung der Richtwirkung festgelegt werden (engl.: Beamforming). In der Literatur werden Beamformer oft als erfolgreich arbeitende Methode gegen Rauschen (z. B. in [BSK01]), teilweise auch bereits gegen Hall beschrieben (z. B. in [OMSG97]). Der Autor erwartet allerdings von DSBs nur eine begrenzt steigernde Wirkung für die Spracherkennung in halligen Umgebungsbedingungen, was nachfolgend begründet wird.

Die Richtcharakteristik kann durch den Bündelungsgrad $\gamma(f)$ (bzw. Bündelungsmaß in dB) beschrieben werden (vgl. Gleichung (3.61)). Für die Problemstellung der vorliegenden Arbeit ist es eine wichtige Tatsache, dass die Richtcharakteristik eine Frequenzabhängigkeit besitzt. Die Kennlinie des Bündelungsmaßes beginnt bei $\gamma(0\ \mathrm{Hz}) = 0$ dB und steigt bei einigen hundert Herz (Beginn der Steigung abhängig von der Geometrie der Mikrofonanordnung) bis hin zum Maximalwert $\gamma_{\max} \approx \mathrm{ld}M \cdot 3$ dB, wobei M der Anzahl der Mikrofone entspricht. Eine Verdopplung von M erzielt demnach eine Steigerung von $\gamma_{\max}$ um 3 dB, was bei einem 4-kanaligen DSB in Hauptrichtung zu einer maximalen Dämpfung um 6 dB von diffusem Rauschen oder Hall führt. Diese Aussagen sind in Abbildung 6.14 simulativ nachgewiesen. Man erkennt, dass beim 4-kanaligen DSB eine sinnvolle Richtwirkung erst bei $f > (1000 \ldots 2000)$ Hz beginnt. Allerdings ist tieffrequenter Hall ($f < 2500$ Hz) laut Abschnitt 4.6.3 gerade besonders störend für die Spracherkennung, höherfrequenter Hall hingegen nicht. Dies und die als eher gering einzuschätzende Dämpfung von 6 dB führen zur Annahme der begrenzten Wirkung der DSBs. Diese Annahme wird durch die Experimente in Abbildung 6.17 bestätigt. Außerdem findet man in [Gar07] Beschreibungen von Experimenten, die im SMART-Room durchgeführt wurden. Es sollten Sprachsignale einer im Fernfeld liegenden Quelle erkannt werden. Dabei wird zwischen Erkennungsraten einer einkanaligen Aufnahme und einer durch einen DSB bearbeiteten Aufnahme verglichen. Als DSB diente das 64-kanalige Mikrofon-Array, das in Abbildung 3.13 angedeutet ist. Die einkanalige Aufnahme wird durch einen Kanal des Arrays zur Verfügung gestellt. Die Erkennungsergebnisse (Tabelle 4.9 und 4.10 in [Gar07]) zeigen kaum nennenswerte Steigerungen durch den DSB (Verringerung der WER von 49,6 % auf 47,4 % bzw. von 40,4 % auf 39,0 %), die den Einsatz eines so aufwendigen Arrays keinesfalls rechtfertigen.

Superdirektive Beamformer [BW01] (SDB) können die Richtwirkung eines DSBs steigern. Allerdings wird diese Steigerung mit neuen praktischen Problemen (instabile Funktionsweise bei ungenauen Angaben von Randbedingungen) erkauft [Hab06]. Bitzer et al. [BSK01] zeigen experimentell, dass Erkennungsraten in halligen Umgebungen

bei Benutzung eines Standard-SDBs gegenüber der Benutzung von DSBs gesteigert werden können.

Adaptive Beamformer Ein Nachteil fixer Beamformer ist, dass die Position des Sprechers bekannt sein muss. Es wurden adaptive Beamformer entwickelt (Ein kurzer Überblick wird in [Hab06] gegeben.), die die Richtung des Schalleinfalls einer Nutzsignalquelle (engl.: Direction of Arrival (DOA)) schätzen und den Beam danach ausrichten. [Hab06] berichtet, dass adaptive Beamformer in halligen Umgebungen nicht zuverlässig arbeiten. Ein typisches Problem ist die Schätzung der DOA, da besonders starke Reflexionen an Wänden einen automatischen DOA-Schätzalgorithmus fiktive Quellen in Richtung dieser Reflexionsstellen vermuten lassen. Der Beam kann folglich fälschlicherweise in Richtung dieser fiktiven Quellen ausgerichtet werden. Adaptives Beamforming ist ein eigenes Forschungsthema. In der vorliegenden Arbeit werden dazu keine weiteren Angaben gemacht.

5.4 Maßnahmen auf Signalebene - blinde Enthallung

Blinde Enthallung von Sprachsignalen ist ein eigenes, noch junges Forschungsgebiet, auf dem erst seit weniger als 10 Jahren Ansätze existieren. Als Anwendung der blinden Enthallung existieren die beiden Möglichkeiten der Verbesserung der Spracherkennung und der Verbesserung der Sprachverständlichkeit von halligen Signalen. Der Raum bzw. das SRM-System wird dabei als unbekannt angenommen, weshalb die Enthallung als blind bezeichnet wird. Da das Thema noch recht jung ist, existieren sehr wenige Artikel (aus den Jahren 2005 – 2008), die einen Überblick über bestehende Ansätze geben, nämlich [NG05], [Hab06] und [HBC08]. In [NJY+08] ist ebenfalls ein Überblicksteil enthalten, der den Stand der Wissenschaft gut abbildet. Interessant ist, dass die einzelnen Artikel sich in der Einteilung und Gewichtung der Maßnahmen teilweise stark unterscheiden, woraus ersichtlich wird, dass sich dieses Forschungsgebiet aktuell noch entwickelt.

Die einzelnen Verfahren lassen sich sowohl nach der Anzahl der benötigten Mikrofone als auch nach den benutzten Prinzipien einteilen. [Hab06] erstellt eine Einteilung nach dem benötigten Grad der Bekanntheit der RIR. Daraus ergeben sich die beiden Klassen Hallunterdrückung (RIR nicht oder wenig bekannt) und Hall-Cancelation[1] (RIR exakt bekannt). Eine Forderung aus Abschnitt 1.2 ist, bei einer Gerätesteuerung mit einem Mikrofon auszukommen. Deshalb soll der Schwerpunkt hier auf einkanaligen Ansätzen liegen, mehrkanalige werden nur kurz erwähnt.

[1] Original verwendet [Hab06] den Begriff Reverberation Cancelation. Eine Übersetzung von Cancelation ins Deutsche fällt jedoch in diesem Zusammenhang schwer, die Möglichkeit (Aus-)Löschung ist ungeeignet und der Begriff Kompensation hat eine nicht zutreffende Bedeutung, vgl. akustische Echokompensation in z. B. [VHH98, Pet01] sowie Geräuschkompensation in [Hau98].

5.4.1 Homomorphic Deconvolution

Die ersten Enthallungsmethoden basieren auf der Eliminierung des SRM-Systems im Cepstralbereich, der die Faltung in (3.65) in eine Addition

$$c_x(q) = c_s(q) + c_h(q) = c_\sim(q) + c_-(q) \tag{5.2}$$

transformiert. Das Prinzip wird in [OS89] unter dem Begriff Homomorphic Deconvolution vorgestellt. Eine Variante ist die Cepstral Mean Subtraction (CMS) von Sprachsignalen [Ata74]. Die Symbole in $c_\sim(q)$ und $c_-(q)$ (5.2) deuten an, dass das Cepstrum des SRM-Systems konstant bleibt (solange sich die SRM-Strecke nicht verändert), die Sprache hingegen zeitlich stark veränderlich ist. Bildet man den Mittelwert von $c_x(q)$, entsteht in etwa $\overline{c}_x(q) \approx c_h(q)$, was zur CMS

$$\hat{c}_s(q) = c_x(q) - \overline{c}_x(q) \tag{5.3}$$

führt. Eine weitere Variante besteht in der cepstralen Kurzpass-Lifterung [BBK91, TLK93], die auf der Annahme beruht, dass Sprache niederquefrent und Hall hochquefrent ist. Einige weitere frühe Enthallungsansätze beruhen auf dem Prinzip von Homomorphic Deconvolution [KR99, PS94, SPW96]. Aufgrund von inkonsistenter Performanz und dem Einbringen neuer Verzerrungen finden jedoch Methoden nach diesem Prinzip für die Enthallung kaum Anwendung, berichtet [HBC08]. Sie gelten später als nicht besonders praktikabel und nicht erfolgreich, wie [EM07, Hab06] schreiben. Dafür werden verschiedene Gründe genannt. Insbesondere können Sprache und Raum in realistischen Szenarien nicht vollständig separiert werden, das Prinzip arbeitet folglich nur in Räumen mit kurzer RIR [EM07]. Weitere Probleme entstehen durch Framingeffekte oder cepstralen Overlap [Hab06].

5.4.2 LP-Residual Enhancement

Eine Klasse von Enthallungsverfahren basiert auf dem Prinzip, Halleffekte im Fehlersignal eines linearen Prädiktors (LP) zu eliminieren (engl.: LP Residual Enhancement). Dabei besteht die Annahme, dass sich nach einer LP-Analyse der Effekt des Halls nicht signifikant in den LP-Koeffizienten widerspiegelt [BG00, HBC08][2]. Im LP-Fehlersignal hingegen wirkt sich der Effekt des Halls aus. Die Idee ist nun, wie in Abbildung 5.1 dargestellt, den Effekt des Halls aus dem Fehlersignal $e_x(t)$ zu eliminieren, um mit dem enthallten Fehlersignal $\hat{e}_s(k)$ eine LP-Synthese durchzuführen, die ein enthalltes Signal $\hat{s}(t)$ ergibt.

[2]Diese Annahme ist in der Praxis nicht nachgewiesen [HBC08]. Sie ist nach den Ausführungen von Kapitel 3 und 4 auch nicht nachvollziehbar. Zusätzlich zeigt [HBC08] graphisch, dass diese Annahme nicht ohne Weiteres getroffen werden kann; zumindest kann sie nicht absolut gelten. [GNW03] zeigt, dass diese Annahme nur in einem räumlich gemittelten Sinn gilt, nicht aber für einen einzelnen Raumpunkt garantiert werden kann [Hab06].

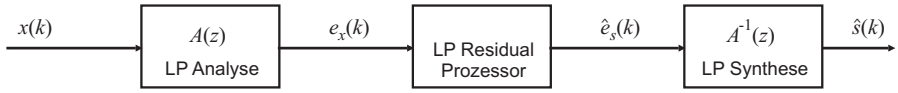

Abbildung 5.1 – Allgemeines Blockschaltbild von Enthallungsmethoden, die auf LP-Residual-Processing beruhen.

Griebel und Brandstein stellen im Jahr 1999 [GB99, BG00] einen Ansatz vor, der mit der Wavelet-Transformation die Maxima in $e_x(t)$ sucht und somit stimmhafte Abschnitte vom Hall bereinigt. Die Methode ist noch unausgereift, sie findet später kaum Anwendung.

Ein wichtigerer Ansatz ist die Methode von Gillespie et al. [GMF01] aus dem Jahr 2001. Sie finden heraus, dass die Gaußähnlichkeit von $e_x(t)$ eine Funktion der Halligkeit ist; je halliger ein Signal ist, um so gaußähnlicher ist $e_x(t)$ und umgekehrt. Die Kurtosis von $e_x(t)$, die ein Maß für die Gaußähnlichkeit darstellt, verhält sich entsprechend (Die Kurtosis beträgt 3 bei vollständiger Gaußähnlichkeit, für $e_x(t)$ von sauberer Sprache ist sie größer als 3 (supergaußisch)). Gillespie et al. schlagen eine adaptive Filterfunktion vor, die $e_x(t)$ enthallen soll, wobei die Maximierung der Kurtosis von $\hat{e}_s(k)$ das Optimierungskriterium der Adaption ist. Erreicht die Adaption eine maximierte Kurtosis, so schließt Gillespie auf ein enthalltes Signal. Die Methode benötigt einige 10 s Sprachsignal zur Adaption[3].

Yegnanarayana und Murthy schlagen im Jahr 1998 einen einkanaligen Ansatz zur Enthallung vor [Yeg98, YM00], in dem sie Kurzzeit-Signalabschnitte von $e_x(t)$ auf einen temporären SRR untersuchen und dabei feststellen, dass dieser aufgrund der stark instationären Eigenschaften von Sprache nicht konstant ist. $e_x(t)$ wird mit einer Wichtungsfunktion modifiziert, die die drei Fälle großer SRR, kleiner SRR und nur Hall unterscheidet.

Yegnanarayana stellt 2002 zusätzlich eine mehrkanalige Methode vor [Yeg02], in der er mit der Hilberttransformation die Einhüllende des gestörten $e_x(t)$ bildet, die dadurch in stimmhaften Abschnitten kräftigere Pulse darstellt. Die rauschartige Hallstörung in $e_x(t)$ wird auf diese Weise verringert. Die Methode kann dadurch eine Verbesserung für Sprachsignale in stimmhaften Abschnitten erreichen [HBC08], allerdings bringt sie auch Störungen in das Signal ein [Hab06].

Gaubitch und Naylor stellen 2004 eine Methode vor [GNW04], die den Ausgang eines Delay-and-Sum-Beamformers mit einer LP-Residual Methode enthallen soll. Dabei bilden sie die Mittelung von angrenzenden Perioden von $e_x(t)$ in stimmhaften Abschnitten, wodurch die nichtperiodischen Anteile (Elemente durch Hallstörungen) unterdrückt bzw. gedämpft werden. Die Methode arbeitet prinzipbedingt nur in stimm-

[3] Quelle: Persönliche Korrespondenz mit K. Kinoshita (vgl. Ansätze in Abschnitt 5.4.3), der die Methode getestet hat.

haften Abschnitten.

Eine weitere Methode, die auf der Enthallung des LP-Fehlersignals basiert, ist die von Delcroix et al. 2006 vorgestellte Methode der Linear Predictive Multi-input Equalization (LIME) [DHM05, DHM07]. Die Methode versucht, das LP-Fehlersignal durch Weißen zu enthallen.

Obwohl die aufgezählten Methoden den Hall verringern können, führen sie u. U. auch zu Signalverzerrungen, die die Natürlichkeit des Signals beeinflussen [NG05].

5.4.3 Spektrale Subtraktion von Hall

Die spektrale Subtraktion ist eigentlich eine Methode, die für die Unterdrückung stationärer Störgeräusche entwickelt wurde (vgl. [Bol79, JH95]). Diese Einschränkung gilt deshalb, weil mit herkömmlichen Verfahren die zu subtrahierenden Störspektren nur von stationären Signalen geschätzt werden können. Wenn man eine Methode findet, die den störenden Hall schätzen kann, obwohl er instationär ist, so ist die spektrale Subtraktion auch als Enthallungsverfahren geeignet. Die folgenden Ansätze schätzen späte Reflexionen und enthallen das Mikrofonsignal durch deren spektrale Subtraktion.

Lebart et al. stellen 2001 [LBD01] eine Methode vor, die späte Reflexionen aufgrund eines statistischen Hallmodells direkt aus dem Mikrofonsignal schätzt und diese spektral subtrahiert. Sie benötigt nur eine Angabe zur Nachhallzeit, die ebenfalls geschätzt wird. Zusätzlich folgt sie der in Kapitel 3 widerlegten Annahme, dass die Nachhallzeit frequenzunabhängig ist.

Wu et al. erweitern die LP-Residual-Enhancement-Methode nach Gillespie (vgl. Abschnitt 5.4.2). Sie entwickeln eine zweistufige Methode, bei der die erste Stufe die von Gillespie vorgeschlagene Maximum Kurtosis Methode ist. Die zweite Stufe schätzt das Leistungsspektrum verbliebener später Reflexionen und führt damit eine spektrale Subtraktion durch. Die Schätzung beruht dabei auf der simplen Annahme, dass später Hall nur eine geglättete und gedämpfte Version von aktuellen Direktschallframes darstellt. Aufgrund dieser Annahme und einigen heuristisch ausgewählten Konstanten, deren Festlegungen im Übrigen sowohl in [WW05, WW06] als auch in seiner Dissertation [Wu03] nicht oder nur ungenügend begründet sind, wird das Spektrum später Reflexionen geschätzt.

Habets schlägt 2005 eine mehrkanalige Methode vor, die späte Reflexionen aufgrund eines statistischen Modells schätzt und spektral subtrahiert [Hab05]. Dabei wird ein räumlich gemitteltes Amplitudenspektrum gebildet, was die Funktionalität eines Delay-and-Sum-Beamformers abbildet und dadurch Hall reduziert. Der geschätzte Hall basiert auf einer Fernfeldannahme, die zur Berechnung eines exponentiellen Abfalls die Nachhallzeit benötigt. Damit dürfte die Methode außerhalb des Fernfelds (Nahfeld) nicht funktionieren. Zusätzlich wird die Nachhallzeit aus einer gegebenen Raumimpulsantwort gemessen. Streng genommen kann die Methode damit nicht mehr als blinde

Enthallung bezeichnet werden.

Die Methode spektrale Subtraktion durch Multi-step Forward Linear Prediction (MSLP) ist eine einkanalige Methode, die von Kinoshita 2006 vorgestellt wird [KNM06, KDNM07]. Unter bestimmten Bedingungen generiert MSLP eine zuverlässige Schätzung für späte Reflexionen [NJY+08]. Deren Schätzung basiert auf einer LP mit einer Verzögerung (Multi-Step Forward LP). [KNM06] demonstriert auch, dass MSLP die Spracherkennung verhallter Signale verbessern kann. Die Methode hat besondere Vorteile gegenüber anderen Ansätzen: Sie arbeitet stabil und benötigt verhältnismäßig geringe Rechenressourcen. Sie kann mit einem Mikrofon betrieben werden und ist zusätzlich robust gegenüber Störgeräuschen [NJY+08]. Laut Kinoshita[4] adaptiert sie sich nach nur 6 s verhalltem Sprachsignal, was gegenüber anderen Methoden der blinden Enthallung derzeit einmalig ist. 2007 veröffentlicht Kinoshita eine mehrkanalige, auf einem ähnlichen Prinzip arbeitende Methode, die neben Hall auch Störgeräusche behandeln kann [KDNM07].

5.4.4 Inversion der Raumimpulsantwort

Eine Reihe von Verfahren verfolgt die naheliegende Idee, die durch die Faltung mit der RIR $h(t)$ eingebrachte Störung durch eine anschließende Faltung mit einem Filter $w_{\text{inv}}(t)$, der das inverse System zu $h(t)$ darstellt, zu eliminieren.

Invertierbarkeit von RIRs Aus der Systemtheorie ist bekannt, dass ein System nur stabil invertiert werden kann, wenn es ein Mindestphasensystem ist. Neely et al. finden allerdings heraus, dass dies bei realen Räumen im Gegensatz zu simulierten Räumen meist nicht zutrifft [NA79].

Hat man nur ein Mikrofon zur Verfügung, ergibt sich die Beziehung

$$\hat{s}(t) = (s * h * w_{\text{inv}})(t) \quad \overset{\mathcal{F}}{\rightarrow} \quad \underline{\hat{S}}(\omega) = \underline{S}(\omega) \cdot \underline{H}(\omega) \cdot \underline{W}_{\text{inv}}(\omega), \tag{5.4}$$

die der Bedingung

$$(h * w_{\text{inv}})(t) = \delta(t) \quad \overset{\mathcal{F}}{\rightarrow} \quad \underline{H}(\omega) \cdot \underline{W}_{\text{inv}}(\omega) = 1 \tag{5.5}$$

genügen muss. Aus der letzten Beziehung ergibt sich $\underline{W}_{\text{inv}}(\omega) = \underline{H}^{-1}(\omega)$. Allerdings ist die RIR aufgrund ihrer Eigenschaft, kein Mindestphasensystem zu sein, zunächst nicht bzw. laut [OS04] nur dann invertierbar, wenn man ein nichtkausales Filter für $w_{\text{inv}}(t)$ benutzt. Das kann praktisch nur erreicht werden, indem die Impulsantwort $w_{\text{inv}}(t)$ um den nichtkausalen Anteil verzögert wird, um wieder ein kausales Filter zu erhalten ($w'_{\text{inv}}(t) = w_{\text{inv}}(t - \tau)$). Dabei entsteht

$$(h * w'_{\text{inv}})(t) = \delta(t - \tau), \tag{5.6}$$

[4] Quelle: Persönliche Korrespondenz mit K. Kinoshita 2007.

sodass das enthallte Signal

$$\hat{s}(t-\tau) = (s * h * w'_{\text{inv}})(t) \tag{5.7}$$

ebenfalls um τ verzögert ist.

Unter der Voraussetzung, dass mehrere Mikrofone benutzt werden können, schlagen Miyoshi et al. [MK88] mit der Methode MINT (engl.: Multiple-Input/Output Inverse Theorem) eine Möglichkeit vor, wie mit einem mehrkanaligen Ansatz der Raum dennoch invertiert werden kann, auch wenn er kein Mindestphasensystem darstellt. Man geht dabei von M Mikrofonen aus, wobei $h_m(t)$ die RIR zwischen Sprecher und dem m-ten Mikrofon darstellt. MINT bildet nun für jeden Kanal ein Entmischsystem $w_{m,\text{inv}}$, das durch die Kombination aller Kanäle ein inverses Filter darstellt. Für die Inversion muss gelten

$$\sum_{m=1}^{M} (h_m * w_{m,\text{inv}})(t) = \delta(t) \qquad \overset{\mathcal{F}}{\rightarrow} \qquad \sum_{m=1}^{M} \underline{H}_m(\omega) \cdot \underline{W}_{m,\text{inv}}(\omega) = 1, \tag{5.8}$$

wobei

$$\hat{s}(t) = \sum_{m=1}^{M} (s * h_m * w_{m,\text{inv}})(t) \qquad \overset{\mathcal{F}}{\rightarrow} \qquad \underline{\hat{S}}(\omega) = \sum_{m=1}^{M} \underline{S}(\omega) \cdot \underline{H}(\omega) \cdot \underline{W}_{m,\text{inv}}(\omega) \tag{5.9}$$

entsteht.

Sowohl für die einkanalige als auch für die mehrkanalige Methode gilt, dass das Originalsignal $s(t)$ perfekt wiederhergestellt werden kann, wenn $h(t)$ bzw. die $h_m(s)$ bekannt sind. Dies ist ein Vorteil der Methode. Ein Nachteil besteht in den enormen Rechenleistungs- und Speicheranforderungen aufgrund der Inversion großer Matrizen. Eine embedded Implementierung in Echtzeit ist daher mit heutigen Hardwaretechnologien als Einprozessorsystem noch nicht realisierbar. Ein weiterer Nachteil ist, dass die Methode sehr sensibel auf Fehlanpassungen der Systeme $h(t)$ bzw. $h_m(t)$ reagiert [HDM07, HBC08]. Die blinde Schätzung der Systeme muss daher sehr genau erfolgen.

Einkanalige Adaption – HERB Nakatani et al. [NM03] stellen 2003 die einkanalige Methode HERB – Harmonicity-based dEReverBeration – vor, die nach dem Prinzip der inversen Filterung nach Gleichung (5.4) arbeitet. Der Algorithmus ist in Abbildung 5.2 als Blockdiagramm dargestellt. Das inverse Enthallungsfilter $w_{\text{inv}}(k)$ wird dabei im Frequenzbereich aus der Division der Spektren des halligen Signals $x(k)$ und eines als sauber angenommenen Signals $x_h(k)$

$$\underline{W}_{\text{inv}}(n) = \mathrm{E}\left\{ \frac{\underline{X}_h(n)}{\underline{X}(n)} \right\} \tag{5.10}$$

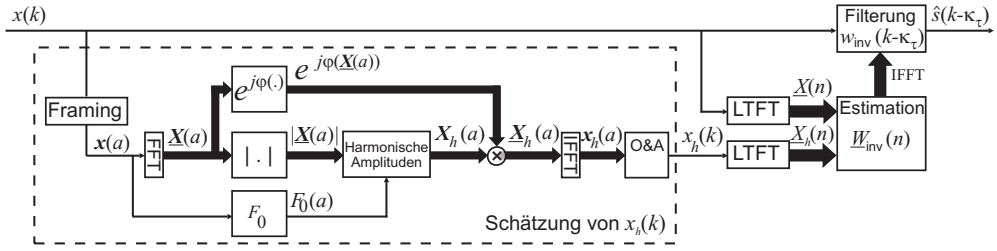

Abbildung 5.2 – Blockschaltbild der Enthallungsmethode HERB.

gewonnen. Da die Spektren von Zeitsignalen einer ganzen Äußerung berechnet werden, wird die dazu benutzte FFT zur Unterscheidung mit dem Begriff Langzeit-Fouriertransformation[5] (LTFT) bezeichnet [NKM07]. Die Mittelung $\mathrm{E}\{.\}$ über mehrere Äußerungen bildet die Schätzung des Filters, das in den Zeitbereich transformiert wird (IFFT $\rightarrow w_{\mathrm{inv}}(k)$). Die o. a. Verschiebung um κ_τ (diskreter Wert für τ) macht das System kausal $w'_{\mathrm{inv}}(k) = w_{\mathrm{inv}}(k - \kappa_\tau)$. Die Faltung mit $x(k)$ ergibt das geschätzte enthallte Signal $\hat{s}(k - \kappa_\tau)$ nach Gleichung (5.7). $x_h(k)$ beinhaltet die harmonischen Komponenten von $x(k)$ und entspricht dem Signal $x_{\mathrm{D},h}(k)$ aus Gleichung (4.10). Es wird als direktes und damit sauberes Signal angenommen. Die Annahme basiert darauf, dass die Harmonischen, wie in Abschnitt 4.3.1 bereits erörtert, als ungestörte Komponenten angenommen werden können, da sie über die Störungen hinausragen. Die gestrichelte Box in Abbildung 5.2 beschreibt die Bildung von $x_h(k)$, indem bei einer Kurzzeitanalyse aufgrund einer F_0-Schätzung ein Spektrum von Harmonischen $\underline{X}_h(n)$ gebildet und dieses in den Zeitbereich zurücktransformiert wird. Die Kurzzeit-Signalabschnitte $x_h(a, k)$ werden mittels eines Overlap-and-Add-Algorithmus (O&A) zu $x_h(k)$ zusammengesetzt. Einzelheiten der Implementierung sind [NMK05, NKM07] zu entnehmen. Problematisch ist, dass laut [NMK05] die Methode HERB ca. eine Stunde Sprachdaten zur ausreichenden Adaption, d. h., zur Schätzung von $w_{\mathrm{inv}}(k)$, benötigt. Kinoshita et al. entwickeln aufgrund dessen mit FastHERB einen schnelleren Schätzalgorithmus, der laut [KNM05] mit nur noch einer Minute Sprachdaten auskommt. Für die Anwendung in der Spracherkennung ist allerdings auch diese Adaptionszeit um Größenordnungen zu groß.

Mehrkanalige Adaption – MINT Wie erwähnt, ist es zwingend nötig, die Systeme $h_m(t)$ genau zu schätzen [HDM07], da sonst erhebliche Fehler bei der Berechnung der Entmischsysteme $w_{m,\mathrm{inv}}(t)$ eingebracht werden, deren Auswirkungen das Signal wesentlich verschlechtern, statt es durch Enthallung zu verbessern. Die genaue Schätzung einer RIR ist eine extrem anspruchsvolle Aufgabe; es handelt sich um ein

[5] Im Unterschied dazu wird die Kurzzeit-Fouriertransformation (STFT) immer auf einen Kurzzeit-Signalabschnitt von z. B. 32 ms angewandt.

FIR-System mit mehreren 1000 Koeffizienten (z. B. 16000 bei 1 s Länge), die blind und nachführend geschätzt werden müssen. Die geforderte Adaptionszeit von weniger als 1 s (vgl. Abschnitt 1.2) stellt neben den hohen Rechenleistungs- und Speicheranforderungen eine zusätzliche, nahezu unhaltbare Randbedingung dar. Obwohl Miyoshi et al. das Konzept von MINT bereits 1988 veröffentlichen [MK88], wird erst 2001 ein Konzept zur blinden Schätzung der $h_m(k)$ entwickelt [Fur01], sodass die MINT-Idee zur blinden Enthallung benutzt werden kann. Dem Finden von Adaptionsstrategien für MINT widmen sich seitdem eine Reihe von Wissenschaftlern, die in den vergangenen fünf Jahren mehrere Lösungsansätze vorgeschlagen haben. Beispiele sind Ansätze von Hikichi et al. 2007 [HDM07] oder Yoshioka et al. [YHM07] (wobei die Methode in [YHM07] auch zu den Ansätzen in Abschnitt 5.4.2 eingeordnet werden kann). In [NJY+08] wird von numerischer Instabilität der Methoden berichtet, wenn die Nachhallzeit steigt, weshalb sie Nachteile in realen Umgebungen haben. In Deutschland befassen sich Buchner, Aichner und Kellermann mit MINT-basierten Enthallungsmethoden [BAK04, BAK05], wobei sie besonders deren algorithmische Verwandschaft zu BSS bzw. BSL Methoden untersuchen.

5.5 TPE-basierte Maßnahmen

Im Unterschied zu herkömmlichen Einteilungen der robustheitssteigernden Methoden für die Spracherkennung bilden in dieser Arbeit die TPE-basierten Maßnahmen eine eigene Gruppe zwischen Ansätzen auf Signalebene und auf Merkmalebene. Begründet wird dies damit, dass TPE-basierte Methoden, die herkömmlich eher den signalbasierten Enthallungsalgorithmen angehören, eigentlich nicht mehr direkt auf dem Signal, sondern eben auf den TPEs arbeiten, die ihrerseits wiederum bereits als Merkmale für die Spracherkennung benutzt werden können (bspw. in [LUA06, LUA08, PLU+08a, PLU+08b]).

Abschnitt 4.2.3 beschreibt, dass Sprache aus zeitlich modulierten Trägersignalen in Subbändern besteht. In Abschnitt 4.3.2 wird dargestellt, dass die Leistungsfunktion der modulierenden zeitlichen Einhüllenden (TPEs) von sauberer Sprache beim Durchqueren eines Raumes ebenfalls verhallt wird und dass dabei die Modulationstiefe sinkt (Abbildung 4.5). Die Übertragungsfunktion des SRM-Systems bezüglich der TPEs ist dabei die MTF (Gleichung (4.30)). Alle bisherigen Methoden basieren darauf, dass die MTF die Form der Fernfeldapproximation nach Schröder (4.41) besitzt. Sie hat Tiefpasscharakter (vgl. Abbildung 4.6), der die TPEs glättet. Houtgast und Steeneken veröffentlichen 1973, dass es einen Zusammenhang zwischen der Modulationstiefe (MTF) und der Sprachverständlichkeit gibt. Mit dem Ziel, die Sprachverständlichkeit zu erhöhen, werden dadurch eine Reihe von Ansätzen motiviert, die versuchen, die verhallten TPEs zu enthallen. Bei der Umsetzung dieses Konzeptes haben sich zwei Herangehensweisen entwickelt:

- **framebasiert** - Die framebasierten Methoden berechnen mit der Kurzzeit-FFT (STFT) zu jedem Zeitpunkt $a = \text{floor}\frac{k}{FI}$ (a - Frameindex, FI - Fortsatzintervall) das Leistungsspektrum. Bei einer bestimmten Frequenz (oder Frequenzband) kann so über zeitliche Fortsetzung der Spektren eine Trajektorie (STPT – Short-Term-Power-Spectrum-Trajectory) beobachtet werden, die den mit der Framerate abgetasteten TPE darstellt. Nach anschließender Enthallung der TPEs wird das Signal wieder mittels Overlap-and-Add [OS04] zusammengesetzt.
- **kontinuierlich** - Die kontinuierlichen[6] Methoden arbeiten im Zeitbereich. TPEs werden aus dem Vollbandzeitsignal bzw. aus den Subbandzeitsignalen einer Filterbank gebildet und stehen mit der vollen Abtastrate f_s zur Verfügung.

Aus der Erkenntnis der Erhöhung der Sprachverständlichkeit wird auch auf die Verbesserung der ASR geschlussfolgert. Bei der Anwendung für die ASR ist nicht die Qualität des Sprachsignals, sondern die der Merkmalvektoren von Bedeutung. Merkmalvektoren werden aus dem (Kurzzeit-)Leistungsspektrum gewonnen. Merkmalvektorfolgen bilden für einzelne Merkmale die von Hermansky beschriebenen zeitlichen Trajektorien. Sie haben somit Verwandschaft zu den TPEs und den STPTs. Deshalb haben, im Vergleich zu signalbasierten Enthallungsmethoden, die TPE- oder STPT-basierten Maßnahmen den Vorteil, dass die Merkmale direkt aus den TPEs bzw. STPTs berechnet werden können und somit das Sprachsignal nicht wiederhergestellt werden muss. Ein weiterer Vorteil sind die daraus ableitbaren, wesentlich geringeren Speicher- und Rechenleistungsanforderungen.

5.5.1 Hochpassfilterung von TPEs

Das Tiefpassverhalten der MTF kann, wie weiter unten beschrieben, durch ihre Inverse (IMTF) kompensiert werden. Dabei beschreibt die IMTF (Gleichung (5.14)) einen Hochpass, der die Einflüsse der MTF ideal aufhebt. Bereits vor den ersten Erkenntnissen zur Arbeit mit der IMTF experimentiert Hirsch (1988) [Hir88] mit Hochpassfilterung von STPTs und erreicht damit in halligen Räumen Steigerungen der Erkennungsrate. Außerdem berichtet Hirsch auch von subjektiv wahrnehmbaren Steigerungen der Qualität von nach der STPT-Hochpassfilterung rekonstruierter Sprache. Diese Ergebnisse waren später Motivation für Arbeiten von Avendano und Hermansky [AH96] (vgl. Abschnitt 5.5.3, Historische Entwicklung). Sie wiederholen das Hochpassexperiment, berichten aber nur von geringen Verbesserungen [AH96].

[6] Der Begriff kontinuierlich ist für abgetastete Signale etwas kritisch, soll aber hier den Unterschied zur Abtastung mit der Framerate symbolisieren, die in den üblichen Systemen der Sprachsignalverarbeitung (je nach f_s und FI) im Verhältnis von 1:80 bis 1:320 zur Abtastrate steht.

5.5.2 RASTA-Filter

Hermansky und Morgan präsentieren 1994 das RASTA-Filter (RelAtive SpecTrAl processing) [HM94b], welches eine Filterung der STPTs darstellt. Das RASTA-Filter ist ein einfaches IIR-Filter und berechnet sich nach

$$H(z) = 0,1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0,98z^{-1}}. \tag{5.11}$$

Es hat sein Passband zwischen (0,26 ... 12,8) Hz (6 dB Abfall vom Durchlassbereich). Die Prinzipien des RASTA-Filters sind im Wesentlichen durch Frequenzen im Modulationsspektrum motiviert (Abschnitt 4.2.3). D. h., sowohl tiefe als auch hohe Modulationsfrequenzen werden als Störung betrachtet und eliminiert. Hier soll, im Gegensatz zum zuvor behandelten Hochpass, besonders die Tiefpassfunktion des RASTA-Filters betont werden. RASTA-Filterung ist eine etablierte Methode zur Steigerung der Robustheit von Spracherkennern [Dro08]. Traditionell handelt es sich dabei um die Robustheit gegen Umgebungsgeräusche. Shire et al. führen im Jahr 1999 auch Experimente zur RASTA-Filterung bei verhallter Sprache durch [SC99, SC00]. Das Ergebnis lautet, dass RASTA-Filter die ASR-Performanz auch in halliger Umgebung steigern können. Der Grund ist, dass RASTA, wie auch schon bei Umgebungsrauschen, die zeitliche Hülle des Signals tiefpassfiltert ($f_{g,o} = 12,8$ Hz). Sie ist durch die Glättung der MTF bereits abgeflacht (Abbildung 4.5, Tiefpassverhalten der MTF). Der steigernde Effekt entsteht hier nicht durch die Entglättung wie bei der Hochpassfilterung, sondern durch die Erhaltung und Betonung der zeitlichen Hüllstruktur der Sprache durch das Eliminieren störender hoher Modulationsfrequenzen.

5.5.3 Enthallung von TPEs mit inverser MTF

Mathematisches Prinzip Mit der Methode der inversen MTF (IMTF) wird versucht, die verhallten TPEs zu enthallen. Gleichung (4.30) beschreibt die Verhallung der TPEs im (Modulations-) Frequenzbereich. Durch Umstellen entsteht der geschätzte enthallte TPE

$$\hat{\underline{M}}_s(\omega_m) = \underline{M}_h^{-1}(\omega_m) \cdot \underline{M}_x(\omega_m) \tag{5.12}$$

mit der inversen Übertragungsfunktion $\underline{M}_h^{-1}(\omega_m)$, die mit der IMTF $m^{-1}(\omega_m)$

$$\frac{1}{W_{\mathrm{e}_h}} \hat{\underline{M}}_s(\omega_m) = m^{-1}(\omega_m) \cdot \underline{M}_x(\omega_m) \tag{5.13}$$

substituiert wird. Mit der Fernfeldapproximation der MTF nach Schröder (4.41) ergibt sie sich zu

$$m^{-1}(\omega_m) = \frac{W_{\mathrm{e}_h}}{|\underline{M}_h(\omega_m)|} = \sqrt{1 + \left(\omega_m \frac{T_{60}}{13,8}\right)^2}. \tag{5.14}$$

Der Proportionalitätsfaktor W_{e_h} sowie die Phase der CMTF in (5.13) werden in den Veröffentlichungen zu diesem Thema normalerweise vernachlässigt, da sie durch unterschiedliche Einstellungen von Mikrofonvorverstärkern bzw. Zeitverzögerungen keine Bedeutung haben und der Fokus auf der Modulationstiefe $m(\omega_m)$ liegt. Die MTF hat Tiefpasscharakter (vgl. Abbildung 4.6), der die TPEs glättet. $m^{-1}(\omega_m)$ hat somit Hochpasscharakter, was zur Wiederherstellung der originalen TPEs führt (Entglätten). Als einzige Unbekannte setzt Gleichung (5.14) die Kenntnis von T_{60} voraus.

Bemerkung: Die IMTF-Methode basiert streng auf der Fernfeldapproximation (4.41). Alle im Folgenden benannten Ansätze verschweigen dies allerdings elegant und damit auch, dass die Methode somit im Nahfeld nicht gültig ist. Dennoch sind sie für das Fernfeld zulässig und erzielen dort mehr oder weniger Wirkung. Vor dem Ansatz von Unoki et al. [UFSA04] (2004) setzen die Verfahren die Kenntnis eines $T_{60,\text{gegeben}}$ voraus und können somit noch nicht als blinde Enthallungsverfahren bezeichnet werden. Weiterhin berücksichtigen sie bis zum Ansatz von Unoki et al. [USFA04][7] nicht die Frequenzabhängigkeit von T_{60}. Daher ergibt sich bei Räumen mit einem frequenzabhängigen $T_{60}(f)$, dass das entsprechende Verfahren für die Frequenzen, bei denen $T_{60}(f) \neq T_{60,\text{gegeben}}$ ist, nicht korrekt arbeitet. Über- oder Unterrestauration sind die Folge (vgl. Gleichung (5.16)). Befindet sich der Sprecher im Nahfeld, wäre die IMTF der Nahfeldapproximation (4.48) zu bilden. Dies ist zum einen mathematisch nicht mehr so einfach zu lösen wie für die Fernfeldapproximation (vgl. Unterschied von (4.48) und (4.41)) und zum anderen werden zusätzlich zum Parameter T_{60} die Werte r_R und r benötigt. D. h., selbst wenn der Raum akustisch vermessen wurde (T_{60} und r_R sind bekannt), spielt trotzdem die Position des Sprechers bzw. der SMD r eine entscheidende Rolle. Für diese Randbedingungen eine *blind* arbeitende Methode zu entwickeln stellt derzeit eine sehr herausfordernde Aufgabe dar.

Historische Entwicklung der IMTF-Methode Der erste Versuch, mit der IMTF die Sprachverständlichkeit zu erhöhen, wird von Langhans und Strube 1982 [LS82] veröffentlicht und arbeitet nach dem framebasierten Prinzip. Die TPEs werden nur von für die Sprachverständlichkeit signifikanten Bändern aus den Kurzzeit-Leistungsspektren gebildet. Langhans und Strube enthallen letztlich nicht die TPEs, sondern das Signal im Spektralbereich

$$\underline{Y}(n,a) = H(n,a) \cdot \underline{X}(n,a). \tag{5.15}$$

Dazu berechnen sie ein Filter $H(n,a)$, welches für jeden Frequenzindex n die Wurzel der IMTF (5.14) enthält und demnach rechnerisch auf die TPEs zurückgeführt werden kann (zur genauen Berechnung von $H(n,a)$ vgl. [LS82]). T_{60} wird als gegeben angenommen, wodurch die Methode noch nicht den Kriterien der blinden Enthallung

[7] Unoki et al. gehen bei ihren Überlegungen von einem frequenzkonstanten T_{60} aus und lösen dieses Problem sozusagen unbewusst. Sie berechnen TPEs in Subbändern und schätzen daraus ein subbandspezifisches T_{60}. Dadurch wird es implizit möglich, ein frequenzvariables T_{60} einzubeziehen.

genügt. Weiterhin berichtet [LS82], dass Experimente zur Sprachverständlichkeit keine Verbesserung bei Benutzung dieser Methode erzielt haben und schlussfolgert deshalb, dass sie nicht für die Enthallung von Sprachsignalen geeignet ist.

Ein zweiter framebasierter Ansatz wird von Avendano und Hermansky 1996 vorgestellt [AH96]. Er arbeitet ähnlich wie die Methode von Langhans und Strube, allerdings wird $H(n, a)$ nicht aus der IMTF gebildet, sondern datenbasiert aus sauberen und verhallten Daten berechnet. Laut [AH96] unterscheidet sich die Übertragungsfunktion des ermittelten Filters im Bereich von 0 Hz $\leq f_m \leq$ 5 Hz nur gering von der IMTF. Oberhalb von 5 Hz wirkt die IMTF als Hochpass, dessen Übertragungsfaktor mit ω_m stetig steigt (5.14); das ermittelte Filter hingegen wirkt oberhalb von 5 Hz als Tiefpass (vgl. Graphik in [AH96]), der Frequenzen oberhalb (5 ... 10) Hz unterdrückt, ein Verhalten, dass dem Prinzip des RASTA-Filters ähnelt. Im Gegensatz zu [LS82] wird in [AH96] von Hörversuchen berichtet, die zwar eine leichte Signalverzerrung beschreiben, aber eine subjektiv wahrnehmbare Verringerung des Halls feststellen.

Als erste kontinuierliche Methode kann der Ansatz von Mourjopoulos und Hammond 1983 [MH83] benannt werden, der Sprache in Subbändern nach Gleichung (4.6) modelliert. Allerdings benutzt er als Träger im Unterschied zu (4.6) ein Kosinussignal. [MH83] leitet weiterhin ab, dass nicht die TPEs nach (4.15), sondern die TMEs gefaltet werden. Die Schlussfolgerung ist eine Enthallung der TMEs in Subbändern durch ein Entfaltungsfilter im Zeitbereich, welches mittels der Wurzel der IMTF berechnet wird. [MH83] berichtet, dass nach Signalrekonstruktion eine subjektiv hörbare Reduktion des Halls in der Sprache beobachtet wird.

Die Methode von Hirobayashi et al. aus dem Jahr 1998 [HNKT98] ähnelt der Methode von Mourjopoulos und Hammond, unterscheidet sich aber durch die Benutzung des Signalmodells (4.6) mit Rauschen als Träger sowie der Annahme der Faltung der TPEs (4.15) anstelle der TMEs. [HNKT98] berichtet, dass die TPE-Enthallung bessere Ergebnisse liefert als die TME-Enthallung.

Praktisch einsetzbare Lösung Unoki et al. [UFSA04, USFA04] stellen im Jahr 2004 eine auf dem Ansatz von Hirobayashi basierende Methode vor, die Gleichung (5.14) umsetzt. Aus Hirobayashis Ansatz leiten sie die folgenden Probleme ab und schlagen entsprechende Lösungen vor:

(i) Demodulation und Ermittelung der TPEs: Amplitudenmodulierte Signale können durch die typischen Techniken wie Gleichrichtung und Tiefpassfilterung [Tay94] oder Synchrondemodulation [Tay94] demoduliert werden. Beide basieren auf einem sinusförmigen Träger. Hier wird jedoch das Modell des Rauschträgers benutzt, sodass Unoki et al. weitere Techniken untersuchen. Insbesondere vergleichen sie die Methode der Tiefpassfilterung des Ensemble Mittels $\langle . \rangle$ mit der Tiefpassfilterung des analytischen Signals (Hüllkurvenberechnung mittels der Hilberttransformation). Letztere wird später sowohl in [UFSA04, USFA04, LUA06, LUA08],

also auch im Rahmen dieser Arbeit (Abschnitt 6.5) sowie in den Veröffentlichungen [PLU+08a, PLU+08b] benutzt.

(ii) Automatische Ermittelung von T_{60}: Durch die automatische (*blinde*) Ermittelung von T_{60} darf Unokis Methode unter den IMTF basierten Ansätzen als erste als *blinde* Enthallungsmethode bezeichnet werden. Prinzipiell wäre es bei einem rein sinusförmigen TPE mit ω_m möglich, den Modulationsgrad $m(\omega_m)$ zu messen und damit T_{60} aus Abbildung 4.6 abzulesen bzw. durch Umstellen von (4.41) zu berechnen. Sprache oder natürliche akustische Signale bestehen jedoch in der Regel aus TPEs, die sich aus mehreren überlagerten Frequenzen zusammensetzen. Unoki et al. entwickeln für solche nicht sinusförmigen TPEs zwei Ansätze zur Ermittelung von T_{60}. Beide basieren darauf, dass es bei der Restauration (Wiederherstellung, Enthallung) der originalen TPEs mit geschätztem $\hat{T}_{60}$ zur neuen Modulationstiefe $m_{\hat{s}}(\omega_m)$ und damit zur

$$\begin{array}{lcll} \hat{T}_{60} < T_{60} & \rightarrow & m_{\hat{s}}(\omega_m) < 1 & \rightarrow \text{ Unterrestauration} \\ \hat{T}_{60} = T_{60} & \rightarrow & m_{\hat{s}}(\omega_m) = 1 & \rightarrow \text{ Optimalrestauration} \\ \hat{T}_{60} < T_{60} & \rightarrow & m_{\hat{s}}(\omega_m) > 1 & \rightarrow \text{ Überrestauration} \end{array} \tag{5.16}$$

kommt. Bei Überrestauration entstehen auch Werte für $\hat{\mathrm{e}}_{\hat{s}}^2(t)$, die sowohl > 1 als auch < 0 sind[8] (vgl. Graphiken in [UFSA04]). Um $\hat{T}_{60,\mathrm{opt}}$ zu ermitteln, wird im ersten Ansatz (vorgestellt in [UFSA04]) der TPE mit (5.13) für verschiedene, größer werdende $\hat{T}_{60}$ im relevanten Bereich $T_{60,\max} \leq \hat{T}_{60} \leq T_{60,\max}$ enthallt. Das größte $\hat{T}_{60}$, bei welchem es keine Abschnitte von $\hat{\mathrm{e}}_{\hat{s}}^2(t)$ gibt, die kleiner als 0 sind, gilt als Messwert für $\hat{T}_{60,\mathrm{opt}}$. Dieses Verhalten kann durch die Vorschrift

$$\hat{T}_{60,c,\mathrm{opt}} = \max\left(\arg \min_{T_{60,\max} \leq \hat{T}_{60} \leq T_{60,\max}} \int_0^\infty \left| \min\left(\hat{\mathrm{e}}_{\hat{s},c}^2(t), 0\right) \right| \mathrm{d}t \right) \tag{5.17}$$

ausgedrückt werden (hier auch für Subbänder, Kanal c). Die zweite Methode zur Bestimmung von $\hat{T}_{60,\mathrm{opt}}$ (vorgestellt in [USFA04]) behandelt das Vorhandensein von längeren Sprachpausen eines TPEs (auch das Ende einer Äußerung ist möglich). Auch hier wird der TPE mit verschiedenen $\hat{T}_{60}$ enthallt. Dabei wird der Pausenzeitpunkt $t_{\mathrm{P},\hat{s}}$ des enthallten TPEs ausgewertet, der durch den Schnittpunkt des abfallenden TPEs mit einem festzulegenden Schwellwert nahe 0 den Beginn einer Pause markiert. $t_{\mathrm{P},\hat{s}}$ wandert bei schrittweiser Erhöhung von $\hat{T}_{60}$ und anschließender TPE-Restauration (Enthallung) zurück in Richtung des Pausenbeginns $t_{\mathrm{P},s}$ des unverhallten Original-TPEs. Wenn $t_{\mathrm{P},\hat{s}}$ den gleichen Wert erreicht wie $t_{\mathrm{P},s}$, dann kann $\hat{T}_{60,\mathrm{opt}}$ angenommen werden. Bei weiterem Erhöhen von $\hat{T}_{60}$ wandert $t_{\mathrm{P},\hat{s}}$ nicht mehr weiter. Dieser Knick in der Funktion $t_{\mathrm{P},\hat{s}}(\hat{T}_{60})$ dient zur

[8] Zu beachten ist, dass es sich bei der Notation $\hat{\mathrm{e}}_{\hat{s}}^2(t)$ nicht um einen quadrierten TME, sondern um einen restaurierten (geschätzten) TPE handelt, der demnach trotz des Quadrates kleiner als 0 sein kann.

Messwertbildung von $\hat{T}_{60,\mathrm{opt}}$. Zur graphischen Erläuterung wird auf [USFA04] verwiesen. Unabhängig von der IMTF-basierten Enthallungsmethode setzen Unoki et al. die Forschungen zur blinden Messung der Nachhallzeit fort und stellen in [UH08a, UH08b] einen weiteren Ansatz vor.

(iii) Gültiges Trägermodell: Die mathematische Ableitung von Gleichung (5.12) geht von dem Rauschträgermodell (4.6) aus. Unoki et al. weisen experimentell nach, dass die IMTF-basierte Enthallung auch für andere Trägersignale, wie sie z. B. auch bei Sprache vorkommen, funktionsfähig ist.

(iv) Bandbreite von Subbändern: Ausgehend von einem Vollbandansatz der IMTF-Methode untersuchen Unoki et al. in [USFA04] Komodulationscharakteristiken von Sprache (ähnliche TPEs benachbarter Frequenzbänder) und ermitteln daraus für eine Subbandmethode die optimale Bandbreite von Filterbankbandpässen mit 100 Hz.

Der IMTF-Ansatz wird durch die Arbeiten von Unoki et al. zu einer praxistauglichen Methode. Die Signalrekonstruktion ist für die Spracherkennung nicht erforderlich. Merkmale können sofort aus den TPEs extrahiert werden. Im Vergleich zu anderen Ansätzen besitzt die Methode relativ geringe Rechenleistungs- und Speicheranforderungen. Eine Adaption gibt es nicht. Allerdings ist ein bestimmter Beobachtungszeitraum nötig, der jedoch nicht länger als die Dauer eines Kommandowortes sein muss (ohne Quelle). Aufgrund dieser zwei wichtigen Vorteile gegenüber anderen Enthallungsmethoden wird angenommen, dass die Methode praktisch einsetzbar und auch für eine embedded Implementierung geeignet ist. Sie wird deshalb mit den im Rahmen dieser Arbeit entwickelten Methoden in Abschnitt 6.5.1.3 auf ihre Leistungsfähigkeit hin verglichen. Außerdem testen bereits Lu, Unoki und Akagi im Jahr 2006 die Methode in Verbindung mit einem Spracherkenner [LUA06, LUA07, LUA08]. Zur genaueren Beschreibung und Einordnung dieser Experimente wird auf Abschnitt 6.5.1.3 verwiesen.

Bemerkungen zur Signalrekonstruktion Da es in der vorliegenden Arbeit um die Verbesserung der Spracherkennung und nicht der Sprachverständlichkeit geht, spielt, wie oben bereits angedeutet, die Signalrekonstruktion aus den TPEs keine Rolle. Der Vollständigkeit halber folgen dennoch einige Gedanken zur Signalrekonstruktion. Dafür wird zusätzlich zum TPE noch der Träger benötigt, auf den bislang nicht eingegangen wurde. Für Houtgast und Steeneken [HS73, HS85] war nur der Modulationsgrad interessant, um die Sprachverständlichkeit vorherzusagen. Drullmann (1994) berichtet ebenfalls von Experimenten, in deren Ergebnis die Struktur der TPEs die wichtigste Signifikanz für die Sprachverständlichkeit besitzt [DFP94a, DFP94b]. Die meisten oben benannten Ansätze folgen diesem Konzept, indem sie TPE-Restauration mit der IMTF vorschlagen. Aufgrund der Ergebnisse von Hörtests wird meist geschlussfolgert, dass die entsprechende Methode für die Verbesserung der Sprachverständlichkeit nicht geeignet ist, wohl aber für einen erfolgreichen Einsatz bei ASR ([LS82, AH96,

HNKT98]). Unoki et al. behaupten, dass die Lösung für die Sprachverständlichkeit in der Rekonstruktion des Originalträgers (zusätzlich zur Enthallung der TPEs) liegt. In [USA03, UTA05] wird gezeigt (2003 und 2005), dass die Sprachverständlichkeit bei TPE-Enthallung tatsächlich gesteigert werden kann, wenn der Träger ebenfalls enthallt wird. Die Trägerrekonstruktion basiert aber in [USA03, UTA05] noch auf einer gegebenen F_0 (in stimmhaften Abschnitten, sonst wird Rauschen benutzt). Offen bleibt noch die Gewinnung eines originalen Trägers aus den verhallten Sprachsignalen. Damit motivieren Unoki et al. ihre späteren Untersuchungen zu F_0-Detektion in verhallten Umgebungsbedingungen. Ausführungen zu diesem Thema finden sich in Kapitel 7 sowie in [UH08c, UHI08, UPMH08, PUM+08], woraus allerdings hervorgeht, dass derzeit eine hinreichend robuste F_0-Detektion unter verhallten Bedingungen noch nicht möglich ist.

5.6 Maßnahmen auf Merkmalebene

Von den zahlreichen Ansätzen auf Merkmalebene zur Steigerung der Robustheit von ASR-Systemen bei Störgeräuschen [Jos02, Dro08] kommen nur wenige für die Störung durch Hall in Frage.

Als erstes wären die dynamischen Merkmale (Δ- und $\Delta\Delta$-Merkmale) zu benennen. Sie sind bereits Bestandteil des UASR-Systems (vgl. Abschnitt 2.2.2). Dynamische Merkmale erhöhen prinzipiell die Robustheit von Spracherkennern, auch unter verhallten Bedingungen, und sind daher keine Methode, die explizit gegen Hall entwickelt wurde.

RASTA-Filter sind bei vielen Autoren der Merkmalebene zugeordnet, da sie neben den TPEs auch auf die Merkmalvektoren angewandt werden können. In dieser Arbeit werden sie bereits in Abschnitt 5.5.2 behandelt. Für das RASTA-Filter gilt ebenfalls, dass es keine Methode ist, die explizit gegen Hall entwickelt wurde. Jedoch steigert es generell, also auch für Hallbedingungen, die Robustheit.

Eine oft erwähnte Methode ist die Cepstral Mean Normalization (CMN) oder Cepstral Mean Subtraction (CMS) (u. a. in [Dro08]). Sie ist für die Kompensation von Störungen gedacht, die durch Kanalverzerrungen auftreten und somit nicht additiver sondern convolutiver Herkunft sind. Im Unterschied zur CMS bei der Homomorphic Deconvolution in Abschnitt 5.4.1 wird hier nicht mit dem Cepstrum des Signals gearbeitet sondern mit MFCC-Merkmalvectoren $\vec{\boldsymbol{x}}_{\mathrm{MFCC}}$. Faltet man zu den Gleichungen (4.1) sowie (4.2), die ein ungestörtes Sprachsignal beschreiben, einen Übertragungskanal $h(t)$, so entsteht

$$\begin{array}{lcccccc} x_{\mathrm{Ph}}(t) & = & e_{\mathrm{Ph}}(t) & * & v_{\mathrm{Ph}}(t) & * & h(t) \\ \underline{X}_{\mathrm{Ph}}(\omega) & = & \underline{E}_{\mathrm{Ph}}(\omega) & \cdot & \underline{V}_{\mathrm{Ph}}(\omega) & \cdot & \underline{H}(\omega). \end{array} \tag{5.18}$$

Mit Bildung des Cepstrums $c(q)$ ergibt sich die Addition

$$c_{x_{\mathrm{Ph}}}(q) \quad = \quad c_{e_{\mathrm{Ph}}}(q) + c_{v_{\mathrm{Ph}}}(q) + c_h(q). \tag{5.19}$$

Wenn eine MFCC-Merkmalanalyse verwendet wird, so ist, wie in Abschnitt 2.2.1 erwähnt, der Anteil des Anregungssignals $e_{\mathrm{Ph}}(t)$ in (5.19) bereits aus dem Merkmalvektor $\vec{\boldsymbol{x}}_{\mathrm{MFCC}}$ eliminiert. Es verbleiben $\vec{\boldsymbol{x}}_{v,\mathrm{MFCC}}$ und $\vec{\boldsymbol{x}}_{h,\mathrm{MFCC}}$, die sich nach zeitlicher Mittelung in einen mittelwertfreien Wert und einen Mittelwert zerlegen lassen

$$\vec{\boldsymbol{x}}_{\mathrm{MFCC}} = \underbrace{\vec{\boldsymbol{x}}_{h\sim,\mathrm{MFCC}} + \vec{\boldsymbol{\mu}}_{h,\mathrm{MFCC}}}_{\vec{\boldsymbol{x}}_{h,\mathrm{MFCC}}} + \underbrace{\vec{\boldsymbol{x}}_{v\sim,\mathrm{MFCC}} + \vec{\boldsymbol{\mu}}_{v,\mathrm{MFCC}}}_{\vec{\boldsymbol{x}}_{v,\mathrm{MFCC}}}. \tag{5.20}$$

Der Vokaltrakt $v_{\mathrm{Ph}}(t)$ ist ein dynamisch veränderliches System, $h(t)$ hingegen ist für die Dauer einer Äußerung nahezu statisch (z. B. Mikrofon), d. h., $\vec{\boldsymbol{x}}_{h\sim,\mathrm{MFCC}} = 0$. Der zeitliche Mittelwert von $\vec{\boldsymbol{x}}_{\mathrm{MFCC}}$ ergibt sich damit zu

$$\vec{\boldsymbol{\mu}}_{x,\mathrm{MFCC}} = \vec{\boldsymbol{\mu}}_{h,\mathrm{MFCC}} + \vec{\boldsymbol{\mu}}_{v,\mathrm{MFCC}}. \tag{5.21}$$

Er wird durch die CMN von $\vec{\boldsymbol{x}}_{\mathrm{MFCC}}$ subtrahiert

$$\hat{\vec{\boldsymbol{x}}}_{\mathrm{MFCC}} = \vec{\boldsymbol{x}}_{\mathrm{MFCC}} - (\vec{\boldsymbol{\mu}}_{h,\mathrm{MFCC}} + \vec{\boldsymbol{\mu}}_{v,\mathrm{MFCC}}) = \vec{\boldsymbol{x}}_{v\sim,\mathrm{MFCC}}. \tag{5.22}$$

Der normierte Vektor $\hat{\vec{\boldsymbol{x}}}_{\mathrm{MFCC}}$ ist von den Anteilen des Kanals befreit. Die Tatsache, dass er zusätzlich von den Mittelwerten der Funktion $\vec{\boldsymbol{x}}_{v,\mathrm{MFCC}}$ befreit ist spielt keine Rolle, da der gleiche Vorgang auch beim Training durchgeführt wird. Der Effekt der CMN liegt im Übereinanderlegen der Gaußverteilungen von Trainings- und Testdaten, die nunmehr den gleichen Mittelwert besitzen. Die Methode kann nur Impulsantworten $h(t)$ kompensieren, die kürzer als die Framelänge sind, was nicht für RIRs zutrifft. Dennoch erreicht man eine Steigerung der Robustheit durch die Normalisierung von Mikrofonen etc.

Sehr et al. [SZK06, SK08] stellen 2006 eine vielversprechende Methode namens REverberation MOdeling for Speech Recognition (REMOS)[9] vor. Sie arbeitet mit einer Kombination von HMMs und einem Hallmodell. Das Hallmodell wird im Wesentlichen durch eine Anzahl von RIRs gebildet, die in der jeweiligen Testumgebung gemessen werden. Es besteht anschaulich aus einer Darstellung der RIRs in Form von Merkmalmatrizen (herkömmliche Merkmalanalyse), aus denen für jedes Matrixelement eine Verteilung ermittelt wird, zu der je ein Mittelwert und eine Varianz gebildet werden. Das HMM bleibt unverändert und wird unverhallt trainiert. Nur der Viterbi-Suchalgorithmus wird durch eine innere Optimierungsprozedur auf das Hallmodell angepasst, d. h., während der Erkennung wird versucht, die Merkmalvektoren von den Hallkomponenten zu befreien, bevor sie auf das unverhallte HMM angewandt werden.

[9] Diesen Namen erhält die Methode erst 2008 in [SK08].

Die Methode stellt daher eine Zwischenlösung zwischen Merkmal- und Modellebene dar. Sie kann derzeit noch nicht als blinde Methode bezeichnet werden, da die Messung der RIRs noch explizit erfolgt, obwohl in aktuellen Experimenten [SWKN08] bereits mit weniger RIRs zur Hallmodellschätzung gearbeitet wurde. Man kann sich jedoch eine adaptive Schätzung der RIR auf Merkmalebene nach Gleichung (3.92) vorstellen, da hier nur noch T_{60} geschätzt werden muss, was wiederum nach der Methode von Unoki et al. (Gleichung (5.17)) durchgeführt werden könnte.

5.7 Maßnahmen auf Modellebene

Die bisher behandelten Methoden verfolgen den Ansatz, die Störung entweder gar nicht erst aufzunehmen oder sie aus dem Signal bzw. den Merkmalvektoren zu eliminieren. Prinzipbedingt ist dies nicht verlustfrei möglich, d. h., die Maßnahmen bringen neben einer Eliminierung der Störung auch Verfälschungen ein. Eine andere Strategie verfolgen die modellbasierten Ansätze. Sie gehen davon aus, dass die Störung bereits im Modell enthalten ist, d. h., ein gestörtes Testmuster wird mit einem gestörten Modell verglichen. Die Abweichung des Testsignals vom Trainingssignal (bzw. Modell) wird somit verringert.

5.7.1 Verhalltes Training

Für additive Geräusche gilt, dass es besonders günstig ist, wenn die Testbedingung bereits im Trainingsmaterial enthalten ist [Dro08]. In [LMP87] werden bereits mehrere Geräuschbedingungen trainiert (Multistyle-Training), um gleich mehrere mögliche Testbedingungen abzudecken.

Auch für Hallstörungen ist ein Training mit verhallten Traingsdaten möglich, um die Robustheit zu steigern. Giuliani et al. [GMOS99] falten die Sprachsignale eines Trainingskorpus mit RIRs einer realen Umgebung und stellen Verbesserungen der Erkennungsrate fest. Auch Stahl et al. berichten in [SFB01], dass die Erkennungsrate in verhallten Umgebungsbedingungen steigt, wenn die Trainingsdaten mit der entsprechenden Bedingung verhallt werden. Haderlein et al. [HNH+05] benutzen zum Training mehrere RIRs eines Raumes bei unterschiedlichen Sprecher-Mikrofon-Positionen. Die Nachhallzeit des Raumes wird in den zwei Stufen $T_{60} = (250;\ 400)$ ms variiert. Für diese eher geringen Nachhallzeiten werden mit dem verhallten Modell gute Erkennungsraten erzielt. Im Rahmen der vorliegenden Arbeit werden u. a. ähnliche Erkennungsexperimente durchgeführt, bei denen zur Ermittlung eines optimalen Trainigs-SMDs verschiedene Sprecher-Mikrofon-Positionen der Testumgebung zum verhallten Training benutzt werden (Abbildung 6.7).

Sehr et al. [SGK06] benutzen anstelle einer RIR für eine bestimmte Trainingsbedingung $T_{60,\mathrm{Train}}$ mehrere RIRs mit unterschiedlichen Nachhallzeiten und bilden daraus

ein Trainingskorpus mit verschiedenen (10) Verhallungen (Multicondition Training[10] (MCT)). Sie zeigen, dass die Erkennungsrate beim MCT für verschiedene Testbedingungen $T_{60,\mathrm{Test}}$ stabiler ist als beim Training mit nur einer Bedingung $T_{60,\mathrm{Train}}$. Im letzteren Fall wird die Erkennungsrate der Testbedingungen, die der Trainingsbedingung entspricht ($T_{60,\mathrm{Test}} = T_{60,\mathrm{Train}}$) besonders bevorzugt (bessere RR als in anderen Fällen). Allerdings erkennt man in [SGK06] auch, dass die maximale Erkennungsrate für MCT (trotz Stabilität) geringer ist als die maximale Erkennungsrate (bei $T_{60,\mathrm{Test}} = T_{60,\mathrm{Train}}$) bei einem zum Vergleich dargestellten Experiment mit verhalltem Training und nur einer Hallbedingung. Als Grund wird vom Autor die durch mehrere Bedingungen entstehende Unschärfe des Modells angenommen. Als Besonderheit kann in [SGK06] die Verwendung von synthetisch erzeugten Impulsantworten angesehen werden. Sie haben allerdings den Nachteil, dass die Nachhallzeit als frequenzkonstant angenommen wird, was in der Praxis, wie bereits mehrfach beschrieben, meist nicht der Fall ist. Außerdem sagt [SGK06] noch nichts über den SMD der synthetischen RIRs aus, obwohl er keine reine Fernfeldapproximation nach Gleichung (4.41) benutzt.

Um den Nachteil des unscharfen Modells zu umgehen besteht die Möglichkeit, mehrere Modelle für je eine unterschiedliche Bedingung zu trainieren. Kann zum Erkennungzeitpunkt die aktuelle Umgebungsbedingung (T_{60} und SMD) bestimmt werden (eventuell adaptiv), so wird das entsprechende Modell zur Erkennung zugeschaltet. Zu einer solchen Vorgehensweise ist dem Autor keine Veröffentlichung bekannt.

Auch in der vorliegenden Arbeit wird das Trainingsmaterial mit verschiedenen Bedingungen verhallt. Die Ergebnisse werden in Abbildung 5.3 dargestellt. Es ist zu erkennen, dass die MFCC-Merkmale für halligere Umgebungen bessere Ergebnisse liefern als die MFB-Merkmale. Dieses Verhalten wird in dieser Arbeit nicht begründet und nicht weiter untersucht, da im Folgenden nur mit den im UASR-System standardmäßig eingestellten MFB-Merkmalen gearbeitet wird. Die MFB-Merkmale dienen als Ausgangspunkt für alle folgenden Experimente. Eine wichtige Erkenntnis ist, dass bei verhalltem Training die Erkennungsraten bei einer Testbedingung, die der Traingsbedingung ähnlich ist, besonders bevorzugt werden (wie schon bei Sehr). Andere Testbedingungen hingegen verschlechtern das Ergebnis, so auch für ungestörte Testdaten. Eine zweite wichtige Erkenntnis ist, dass ein verhalltes Training bei der Fernfeldbedingung ungünstigere Ergebnisse für alle Testbedingungen generiert. Es wird daher vermutet, dass trotz Hall ein wesentlicher Direktschallanteil im Trainingsmaterial enthalten sein muss, damit signifikante Sprachmerkmale in das Training des Modells einfließen können. Aufgrund dieser Vermutung wird später die Abhängigkeit vom Trainings-SMD in Abbildung 6.7 noch einmal detaillierter untersucht. Dabei wird die Schlussfolgerung gezogen, dass ein SMD, der kleiner als der Hallradius ist, aber dennoch in dessen Nähe liegt, besonders günstig ist.

[10] Es handelt sich dabei nur um einen anderen Begriff für das oben bereits angesprochene Multistyle-Training. Beide Begriffe werden in den entsprechenden Veröffentlichungen definiert, haben aber die gleiche Bedeutung.

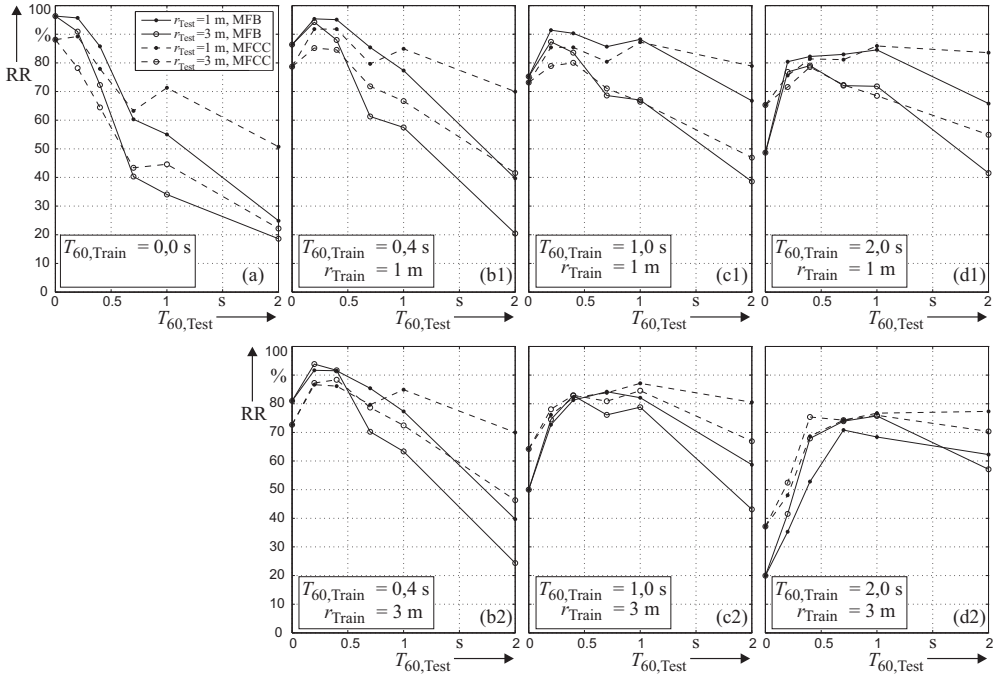

Abbildung 5.3 – Erkennungsrate bei verhalltem Training für MFB- und MFCC-Merkmalvektoren. (a) saubere Trainingsbedingung, (b1) ... (d2) verhalltes Training $T_{60,\mathrm{Train}} = (0,4;\ 1,0;\ 2,0)$ s, SMD = (1; 3) m (oben; unten).

Alle zitierten Veröffentlichungen lassen eine systematische Untersuchung der Abhängigkeit der Erkennungsrate von der Nachhallzeit *und* vom SMD, wie sie in Abbildung 5.3 dargestellt ist, aus. Wenn überhaupt eine Abhängigkeit vom Grad der Verhallung dargestellt ist, wird für gewöhnlich nur eine Abhängigkeit von T_{60} ermittelt, wobei teilweise noch nicht einmal eine Angabe zum benutzten SMD gegeben ist. Deshalb stellt die systematische Untersuchung nach diesen beiden Parametern in dieser Arbeit eine Neuheit dar.

5.7.2 Modelladaption

Wie erwähnt ist es nicht möglich, für sämtliche vorkommenden Testbedingungen ein Modell zu trainieren. Deshalb wurden für den Fall von additiven Störgeräuschen verschiedene Modelladaptionstechniken entwickelt (z. B. PMC[11] [Gal95], VTS[12] [MRS96],

[11] Engl: PMC – Parallel Model Combination.
[12] Engl: VTS – Vector Taylor Series.

HMM Decomposition [VM90]). Sie benutzen ein sauber trainiertes Modell, um anschließend verschiedene Parameter zu adaptieren, sodass sich das Modell an die Störung anpasst. Dabei basieren sie auf der Annahme, dass sich das Geräusch additiv im Merkmalvektor auswirkt. Deshalb sind diese Techniken zunächst nicht für Störungen durch Hall geeignet [SK08]. Raut et al. [RNS06] sowie Hirsch et al. [HF06] entwickeln zwei Modelladaptionsansätze, die jeweils für lange Nachhallzeiten konzipiert sind. Sie benötigen T_{60} als Eingangsparameter und basieren auf einem Fernfeldmodell. Dabei wird aus einem aktuellen HMM-Zustand auf Folgezustände geschlossen und deren Mittelwerte entsprechend angepasst. In beiden Veröffentlichungen wird keine Aussage zum SMD gemacht, beide zeigen auch nur wenige Evaluationsergebnisse, sodass über die Funktionsfähigkeit der Methoden hier keine zuverlässige Aussage getroffen werden kann.

5.8 Zusammenfassung der Erkenntnisse

In diesem Kapitel wurde versucht, den Stand der Technik darzustellen, den die Methoden beschreiben, die zur Robustheitssteigerung von Spracherkennern unter Hallbedingungen beitragen können. Eine Überblicksdarstellung ist vergleichsweise neu, es existieren kaum Artikel zu diesem Thema (z. B. [SK08]). Die Einteilung der Methoden lehnt sich an die Einteilung zu Maßnahmen zur Robustheitssteigerung von ASR-Systemen gegen Störgeräusche an. Es existieren Methoden auf der akustischen, auf Signal-, Merkmal- sowie Modellebene. Zusätzlich zu dieser klassischen Einteilung werden TPE-basierte Maßnahmen als eine eigene Gruppe zwischen Signal- und Merkmalebene aufgeführt. Laut [Dro08] darf angenommen werden, dass das Training mit gestörten (verhallten) Daten, die die Testbedingung abbilden, die leistungsfähigste Methode darstellt. Die Experimente in Kapitel 6 werden diese Annahme belegen. Verhalltes Training hat weiterhin den Vorteil, dass es keinerlei Adaptionszeit sowie zusätzliche Rechenressourcen benötigt. Ein Nachteil ist allerdings, dass die Testbedingung zum Trainingszeitpunkt bekannt sein muss. Allerdings kann dies, wie in Kapitel 3 ermittelt wurde, für Räume im Wohn- und Büroumfeld zumindest in den Bereich $T_{60} = (0,3 \ \ldots \ 0,8)$ s und $r = (0,5 \ \ldots \ 4)$ m eingeschränkt werden. Es besteht nun die Aufgabe, eine geeignete Trainingsbedingung zu finden, die diesen Bereich ausreichend abdeckt. Das verhallte Training kann durch weitere Methoden unterstützt werden. Die meisten Enthallungsmethoden eignen sich aufgrund der Randbedingung allerdings nicht für den praktischen Einsatz. Dies betrifft insbesondere die geforderte geringe Adaptionszeit, die teilweise enormen Rechenleistungs- und Speicheranforderungen sowie die eingeschränkte Leistungsfähigkeit der Methoden für reale Bedingungen. Aus den existierenden Ansätzen wurden zum Vergleich die Methoden TPEFA, IMTF-basierte Enthallung sowie DSB herangezogen (vgl. Abschnitt 6.5), da sie zumindest in den Punkten Adaptionszeit sowie Speicher- und Rechenleistungsanforderungen im praktisch einsetzbaren Bereich liegen.

6 Harmonicity-based Feature Analysis

6.1 Überblick

In Kapitel 5 werden bereits veröffentlichte Ansätze vorgestellt, die zur Verbesserung der Spracherkennung unter raumakustischen Umgebungen in Betracht kommen. Keiner der bestehenden Ansätze erfüllt vollständig die für den Praxiseinsatz in einem sprachgesteuerten Gerät im Wohn- und Büroumfeld erforderlichen Anforderungen:

- Steigerung der Erkennungsrate auf RR > 90 %,
- Robustheit der Erkennungsrate gegen variierende Umgebungsbedingungen (variierende Räume (T_{60}) sowie variierende Sprecherpositionen (SMD)),
- Abdeckung typischer Wohn- und Bürobedingungen: $T_{60} = (0,3 \ \ldots \ 1,0)$ s; $r_{\text{Test}} = (0,5 \ \ldots \ 4)$ m,
- keine bzw. geringe Adaptionszeit (< 1 s),
- geringe oder moderate Anforderungen an Rechenleistung und Speicherbedarf, da in Geräten normalerweise eine eingebettete Implementierung benötigt wird.

In diesem Kapitel wird ein ein neuartiges Verfahren entwickelt, das bereits im Vorfeld an diesen Anforderungen ausgerichtet ist. Inspirationen zu diesem Verfahren basieren auf der Enthallungsmethode (HERB – Harmonicity-based Dereverberation) von Nakatani et al. [NKM07] (vgl. Abschnitt 5.4.4). Das hier vorgestellte Verfahren arbeitet ebenfalls mit dem Prinzip von harmonischen Komponenten der Sprache, weshalb der recht ähnliche Begriff Harmonicity-based Feature Analysis (HFA) gewählt wird. Im Unterschied zu HERB handelt es sich nicht um eine Enthallungsmethode, sondern, wie der Name andeutet, um eine Maßnahme der robusten Signal- und Merkmalanalyse für die Spracherkennung unter raumakustischen Umgebungsbedingungen.

Zunächst wird das Konzept der Methode vorgestellt (Abschnitt 6.2). Es basiert auf drei Ideen, die zum Teil auf in vorhergehenden Kapiteln erarbeiteten Erkenntnissen beruhen. Im Anschluss wird der Algorithmus von HFA vorgestellt, der diese drei Ideen umsetzt (Abschnitt 6.3). Er gliedert sich in vier Submodule, von denen das Kernstück eine spektrale Synthese ist.

Im Anschluss folgen Experimente, die das Verhalten von HFA mit einer herkömmlichen Merkmalanalyse vergleichen. Die Experimente erfolgen sowohl mit unverhalltem als auch mit verhalltem Training. Die durch Raumhall bestehenden Abhängigkeiten werden dabei umfassend untersucht. Der Vorteil gegenüber der herkömmlichen Merkmalanalyse wird deutlich herausgestellt.

Um die hier entwickelte Methode HFA mit bestehenden Verfahren zu vergleichen, werden zwei weitere Methoden implementiert und teilweise auch mit HFA kombiniert. Dies ist zum einen die TPE-Feature-Analysis (TPEFA), die auf der Extraktion von TPEs beruht, aus denen durch einen Tiefpass störende hohe Modulationsfrequenzen eliminiert werden (vgl. Abschnitt 6.5.1.1). Die Methode wird in dieser Arbeit weiterentwickelt. Zum anderen wird eine Enthallungsmethode der TPEs durch die inverse MTF (IMTF) getestet (vgl. Abschnitt 6.5.1.3). Zusätzlich wird HFA mit der Leistungsfähigkeit eines Delay-and-Sum Beamformers (DSB) verglichen. Weitere Methoden werden nicht untersucht, was u. a. daran liegt, dass derzeit noch kein Ansatz existiert, der das Problem im Rahmen der oben angesprochenen Randbedingungen für den praktischen Einsatz zufriedenstellend löst.

In vergleichenden Experimenten liefert TPEFA gute Ergebnisse. Es wird festgestellt, dass HFA und TPEFA sich jeweils ergänzen und in einer kombinierten Variante (Abschnitt 6.5.1.2) zu bemerkenswerten Erkennungsraten führen, insbesondere bei verhalltem Training. Das Verhalten der IMTF-Methode wird ebenfalls untersucht. Es stellt sich wie erwartet heraus, dass sie nur im Fernfeld arbeitet. Dort liefert sie die besten Ergebnisse, die allerdings mit denen der anderen Ansätze bei verhalltem Training nicht konkurrieren können. Im abschließenden Vergleich mit dem DSB schneidet HFA deutlich besser ab. Es kann gezeigt werden, dass DSBs wie erwartet für die Hallproblematik keine Lösung sind; sie können jedoch unterstützend wirken. Aus den Experimenten ist zu entnehmen, dass HFA einen technischen Fortschritt bietet, teilweise besser arbeitet als existierende Verfahren und diese durch HFA hervorragend ergänzt werden können. Durch HFA werden die oben genannten Kriterien für den Praxiseinsatz in einem sprachgesteuerten Gerät im Wohn- und Büroumfeld erfüllt.

6.2 Konzeption von HFA

Die Methode HFA wurde entwickelt, um die stark einbrechenden Erkennungsraten in halligen Umgebungen zu verbessern (Abbildung 4.9). Sie steigert die Robustheit der Spracherkennung. Dabei wird eine Folge von Spektren generiert, die auf sicheren extrahierbaren Informationen im Sprachsignal basieren. Merkmale, die umgebungsspezifisch sind, sollen dabei möglichst ausgelassen werden. Trotz variierender Umgebungsbedingungen entstehen so ähnliche Merkmalmuster, die zu besseren Erkennungsergebnissen führen. Bei dieser Vorgehensweise gehen Informationen verloren, die normalerweise für die Spracherkennung benutzt werden. Dieser (geringe) Verlust wird zu Gunsten einer unterdrückten (starken) Störung in Kauf genommen. Die Gewinnung sicherer Merkmale basiert auf den folgenden drei Ideen:

(i) Harmonische Komponenten werden als ungestört angenommen: Die Idee von HFA ist durch die Arbeit von Nakatani et al. [NKM07] inspiriert worden. Sein Ansatz HERB (Harmonicity-based dEReverBeration) versucht ein zur RIR inverses System zu schätzen und damit das verhallte Signal zu enthallen (vgl. auch Abschnitt 5.4.4). Die Schätzung des Filters erfolgt durch Gleichung (5.10), die das als ungestört angenommene Signal $\hat{x}_{\mathrm{D},h}(t)$ benötigt, das aus in Kurzzeitabschnitten gemessenen harmonischen Komponenten $x_{\mathrm{D},h}(t-\tau)$ (vgl. Gleichung (4.11)) synthetisiert wird. Deren Maxima schauen aus dem Spektrum heraus und heben sich somit von einem flacheren Rausch- oder Hallteppich ab (vgl. auch Abbildung 4.4 Markierung v), wodurch sie als ungestört betrachtet werden können. Im Gegensatz zu HERB greift HFA bereits vor der Synthese und der Berechnung eines Enthallungsfilters ein und misst die (als ungestört angenommenen) harmonischen Komponenten $x_{\mathrm{D},h}(t-\tau)$, um daraus eine Folge von Spektren aus Harmonischen zu generieren (spektrale Synthese). Daraus werden Merkmalvektoren gewonnen, die als ungestört angenommen werden. Der Erkenner wird im Vorfeld bereits mit ebenso analysiertem Trainingsmaterial trainiert. Demnach umgeht HFA die sensible und zugleich Zeit konsumierende Komponente der Filterschätzung. Es verbleiben die Komponenten der F_0-Detektion und der spektralen Synthese. Eine Adaptionszeit ist nicht nötig.

(ii) Im unteren Frequenzbereich sind stimmlose Laute durch Hall vorhergehender stimmhafter Laute stark gestört: Saubere stimmlose Laute, z. B. Frikative, sind typischerweise breitbandig, wobei sich die wichtigsten Merkmale meist im oberen Frequenzbereich befinden. Wie in Abschnitt 4.3.1 beschrieben wurde, sind die Bereiche im Frequenzband unter $\approx$ 2900 Hz in stimmlosen Regionen stark durch die hineinhallenden Formanten eines vorhergehenden stimmhaften Lautes gestört. Die störenden Frequenzen sind dabei durch die energiereichen Frequenzbänder des stimmhaften Lautes bestimmt. Durch die starke Energie stimmhaften Halls, die zu signifikanten Ausprägungen tieffrequenter Merkmale in der spektralen Hülle des gestörten stimmlosen Lautes führt, weichen die Spektren von gestörten und ungestörten stimmlosen Lauten voneinander ab. Ein sauber trainiertes Phonemmodell für einen stimmlosen Laut würde also bei Anlegen eines gestörten stimmlosen Lautes eine geringe Emissionswahrscheinlichkeit erzeugen; eventuell würde ein stimmhaftes Phonemmodell sogar eine größere Wahrscheinlichkeit liefern. Um sowohl die spektrale Struktur von ungestörten als auch von gestörten stimmlosen Lauten einander ähnlich zu gestalten, werden in HFA tiefe Frequenzen sowohl im gestörten als auch im ungestörten Fall unterdrückt. Im Ergebnis werden nur noch die hochfrequenten Merkmale im Signal belassen bzw. zu einer spektralen Synthese hinzugezogen. Im Falle von breitbandigen stimmlosen Lauten, wie Plosiven, gehen damit tieffrequente Nutzinformationen verloren. Da dies auch für das Training gilt, entsteht nur ein geringer Nachteil.

(iii) Hochfrequenter Hall ist für die Spracherkennung harmlos: Die Experimente in Abbildung 4.11 (d) zeigen deutlich, dass sich hochfrequenter Hall über 2500 Hz nur gering bis gar nicht störend auf die Spracherkennung auswirkt. Abbildung 4.11 (c) zeigt sogar einen leicht positiv wirkenden Effekt hochfrequenten Halls. Diese Eigenschaften des hochfrequenten Halls werden in HFA genutzt. Die zwei zuvor beschriebenen Ideen, **(i)** für stimmhafte Abschnitte und **(ii)** für stimmlose Abschnitte, werden nur im tieffrequenten Bereich (z. B. < 2900 Hz) umgesetzt. Darüber (z. B. > 2900 Hz) wird das Signal nicht verändert, es bleibt original erhalten.

Wie bereits angedeutet, ist es erforderlich, den Erkenner mit Merkmalen zu trainieren, die durch den gleichen Analyseprozess von HFA generiert werden.

6.3 Algorithmus

HFA wird als Modul in die Signalverarbeitungskette eines Merkmalanalysators zwischen FFT und Mel-Filterbank eingefügt (vgl. Abbildung 6.1). Der benutzte Merkmalanalysator basiert auf der im UASR-System verbauten Merkmalanalyse, die in Abschnitt 2.2.1 kurz beschrieben wird. Er wird in dieser Arbeit mit 'herkömmliche Merkmalanalyse' (engl.: Conventional Feature Analysis – CFA) bezeichnet. HFA erhält am Eingang die FFT-Betragsspektren $|\underline{X}(a,n)|$ im positiven Frequenzbereich. Am Ausgang erzeugt es die synthetischen Spektren $X_S(a,n)$, aus denen mit der Mel-Filterbank die Merkmalvektoren $\vec{\boldsymbol{x}}_{\mathrm{MFB}}$ extrahiert werden. HFA besteht aus den vier Submodulen:

- F_0-Detektion,
- Bildung des harmonischen Amplitudenspektrums,
- VUD – Stimmhaft-Stimmlos-Entscheidung und
- spektrale Synthese.

Sie werden im Folgenden beschrieben.

6.3.1 F_0-Detektion

Sowohl für das nachfolgend beschriebene generierte Spektrum von Harmonischen als auch für die VUD wird als Eingangsinformation die Grundfrequenz F_0 des Sprachsignals benötigt. Aus diesem Grunde beschreibt Abschnitt 7.4 mehrere F_0-Detektionsverfahren und untersucht diese unter halligen Bedingungen auf ihre Funktionalität hin. Im Ergebnis kann festgestellt werden, dass keines der Verfahren unter Hallbedingungen korrekt arbeitet. Allerdings werden die besten fünf Verfahren benannt, von denen keines signifikant besser arbeitet als die anderen vier (vgl. Abschnitt 7.4.2). Deshalb wird in dieser Arbeit der relativ simple Ansatz der ACF (vgl. Abschnitt 7.4.1) gewählt,

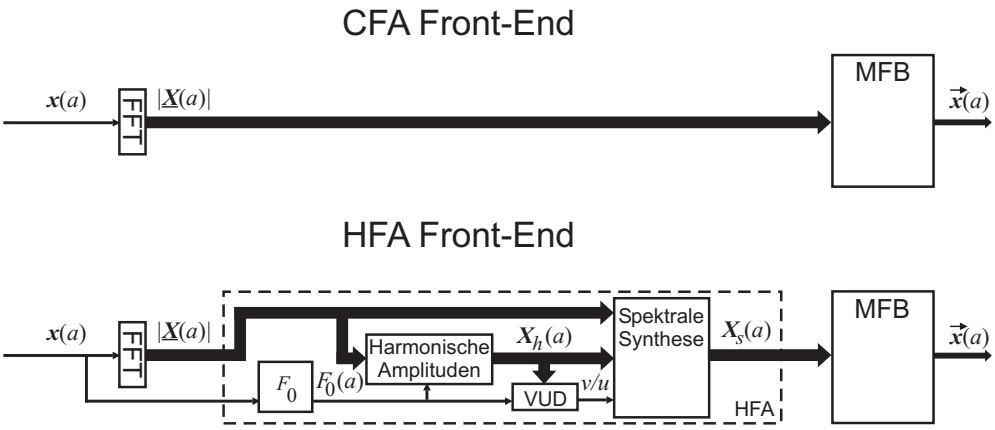

Abbildung 6.1 – Oben: CFA – Conventional Feature Analysis, herkömmliche Merkmalanalyse, bestehend aus FFT und Mel-Filterbank im Spektralbereich. Unten: HFA – Harmonicity-based Feature Analysis, besteht zusätzlich zu den CFA-Komponenten aus F_0-Detektion, VUD, Detektion harmonischer Amplituden und spektraler Synthese. Sowohl Zeit- als auch Frequenzsignale werden in Blockschreibweise (fett) dargestellt, die die frameinternen Indizes k und n weglassen.

gefolgt von einem Postprocessingverfahren der Impulsunterdrückung (vgl. Abschnitt 7.5.2). Beide benötigen nur geringe Rechenleistung. Im Unterschied zur CFA benötigt HFA deshalb noch die Matrix des in Frames zerlegten Zeitsignals $x(a,k)$ (Abbildung 6.1) als Eingangssignal, aus der frameweise die ACF berechnet wird. Um Rechenleistung zu sparen, wird die ACF in der Implementierung aufgrund der Maximumsuche nur in den Grenzen $T_{0,\min}$ und $T_{0,\max}$ berechnet. Die entsprechenden Limits aus (4.5) ergeben

$$\begin{array}{rcrcrcr} \kappa_{0,\min} & = & \mathrm{rd}(T_{0,\min} \cdot f_s) & = & \mathrm{rd}(\quad 1,67\ \mathrm{ms} \cdot 16\ \mathrm{kHz}\ ) & = & 27 \\ \kappa_{0,\max} & = & \mathrm{rd}(T_{0,\max} \cdot f_s) & = & \mathrm{rd}(\quad 14,28\ \mathrm{ms} \cdot 16\ \mathrm{kHz}\ ) & = & 228. \end{array} \tag{6.1}$$

Der Rundungsoperator rd(.) rundet dabei zur nächst liegenden natürlichen Zahl. Mit diesen Werten ergibt Gleichung (7.11) die implementierte ACF

$$\psi_{xx}(a,\kappa) = \sum_{k=0}^{K-\kappa-1} x(k) \cdot x(k+\kappa)\ ; \quad \kappa_{0,\min} \leq \kappa \leq\ \kappa_{0,\max}, \tag{6.2}$$

gefolgt von der Maximumsuche nach Gleichung (7.10)

$$\kappa_0(a) = \arg \max_{\kappa_{0,\min}}^{\kappa_{0,\max}} \psi_{xx}(a,\kappa), \tag{6.3}$$

womit die vorläufige Grundfrequenz

$$F_0'(a) = \frac{f_s}{\kappa_0(a)} \tag{6.4}$$

berechnet wird. Auf die Sequenz $F_0'(a)$ wird noch, wie erwähnt, das Postprocessingverfahren der Impulsunterdrückung angewendet, welches letztlich den Grundfrequenzverlauf $F_0(a)$ ergibt. $F_0(a)$ enthält für jeden Frame einen F_0-Wert, so auch für stimmlose Frames, die ja eigentlich kein F_0 besitzen. Dies wird mit der anschließenden VUD korrigiert. Für Einzelheiten der F_0-Detektion wird auf Kapitel 7 verwiesen.

6.3.2 Bildung des harmonischen Amplitudenspektrums

Um ein für die spektrale Synthese benötigtes Spektrum von Harmonischen $X_h(a,n)$ zu berechnen, wird $|\underline{X}(a,n)|$ an den spektralen Indizes der Harmonischen $n_{h,i}(a)$ gemessen. Diese Indizes werden am i-ten Vielfachen der Grundfrequenz berechnet

$$n_{h,i}(a) = \mathrm{rd}\left(\frac{i \cdot F_0(a)}{\Delta f}\right) \;\; ; \quad i = 1,\ \ldots,\ i_{\max}(a)\ , \tag{6.5}$$

wobei aufgrund der framespezifischen Grundfrequenz $F_0(a)$ die framespezifische Obergrenze

$$i_{\max}(a) = \mathrm{rd}\left(\frac{f_{\max}}{F_0(a)}\right) \tag{6.6}$$

berechnet wird, die zur maximalen einbezogenen Frequenz $f_{\max}$ gehört. Dabei wird $f_{\max}$ mit 6000 Hz so eingestellt (heuristisch), dass möglichst die höchsten signifikanten Harmonischen einbezogen werden. $f_{\max}$ kann variiert werden, dies wurde in dieser Arbeit jedoch nicht untersucht. Die Indizes der Harmonischen führen zum (Amplituden-) Spektrum der Harmonischen

$$X_h(a,n) = \begin{cases} |\underline{X}(a,n)| & ; \quad n \in n_{h,i}(a) \\ 0 & ; \quad n \notin n_{h,i}(a) \end{cases} \tag{6.7}$$

mit $n_{h,i}(a)$ nach Gleichung (6.5). $X_h(a,n)$ stellt in stimmhaften Frames eine Abtastung von $|\underline{X}(a,n)|$ an den Maxima der harmonischen Komponenten dar (vgl. Abbildung 6.3, $X_h(a,n)$ wird durch die dargestellten Spektrallinien markiert). In stimmlosen Frames stellt es die Abtastung des stimmlosen Spektrums, welches bspw. ein Rauschspektrum sein kann, bei Vielfachen eines fiktiven F_0 dar. Eine Unterscheidung in stimmhafte und stimmlose Frames ist erst durch den nachfolgenden VUD möglich, der $X_h(a,n)$ als Eingangssignal erhält.

6.3.3 VUD – Stimmhaft-Stimmlos-Entscheidung

Da die spektrale Synthese von HFA stimmhafte und stimmlose Frames unterschiedlich behandelt, wird ein VUD benötigt, der diese Unterscheidung vornimmt. Deshalb werden in Abschnitt 7.6 VUD-Verfahren beleuchtet. Ein schwellenbasierter VUD wurde implementiert und unter Hallbedingungen untersucht. Motivation, Einzelheiten der Implementierung sowie Vor- und Nachteile des gewählten Ansatzes sind in Abschnitt 7.6.2 beschrieben. Eine umfassende Evaluation des Ansatzes findet sich in Abschnitt 7.6.3. Dabei spielt insbesondere die Einstellung des Schwellwertparameters b_{th} (bzw. $L_{b_{th}} = 10 \lg b_{th}$), der in Gleichung (7.39) benötigt wird, für die Funktion des VUD eine besondere Rolle. Bei sauberen Sprachdaten erweist sich $L_{b_{th}} = (-10 \ldots -5)$ dB als besonders günstig; bereits bei geringer Verhallung ergibt sich ein optimales $L_{b_{th}}$ bei -2 dB, tendierend zu 0 dB bei stärkerer Verhallung. Das Verfahren wird außerdem in Zusammenarbeit mit HFA für die Spracherkennung auf seine Funktionalität hin überprüft (vgl. Abschnitt 6.4.3). Auch hier spielt die Einstellung des Parameters b_{th} eine wichtige Rolle. Es zeigt sich, dass $L_{b_{th}} = 0$ dB eine günstige Einstellung darstellt. Letztlich ist zu erwähnen, dass dieser recht einfache Ansatz für die Methode HFA ausreichend ist; eine begrenzte Anzahl an Ungenauigkeiten in der VUD-Entscheidung wird bei HFA einkalkuliert und ist eventuell sogar gewünscht (vgl. dazu Abschnitt 6.3.6).

6.3.4 Spektrale Synthese

Die spektrale Synthese ist das Kernmodul von HFA, da es die drei zentralen Ideen **(i)**, **(ii)** und **(iii)** aus Abschnitt 6.2 umsetzt. Sie generiert, wie der Name sagt, ein synthetisches Spektrum $X_S(a, n)$, wobei die Synthese für stimmhafte (Index v wie engl.: voiced) und stimmlose (Index u wie engl.: unvoiced) Frames unterschiedlich erfolgt. Deshalb wird festgestellt, dass zwei unterschiedliche Analysemethoden existieren, eine für stimmhafte und eine für stimmlose Frames.

Da die spektrale Synthese nicht frameübergreifend arbeitet, wird der Frameindex a aus Vereinfachungsgründen innerhalb dieses Unterabschnitts weggelassen.

6.3.4.1 Stimmlose Frames

Für Frames, die der VUD als stimmlos klassifiziert, werden die Ideen **(ii)** und **(iii)** aus Abschnitt 6.2 implementiert. Das bedeutet, hochfrequente Anteile werden unverändert belassen (Idee **(iii)**) und tieffrequente Anteile werden gelöscht (Idee **(ii)**). Das Löschen ist kein hartes Ersetzen mit Null, sondern es wird ein energiearmer Geräuschboden $X_{Fl}(n)$ eingefügt. Dafür gibt es zwei Gründe. Zum einen würde der sich anschließende Logarithmus in der Melfilterbank den Wert Null in $-\infty$ umwandeln[1]. Zum anderen

[1] Es ist deshalb ohnehin sinnvoll, für den Logarithmus einen Minimalwert zu definieren.

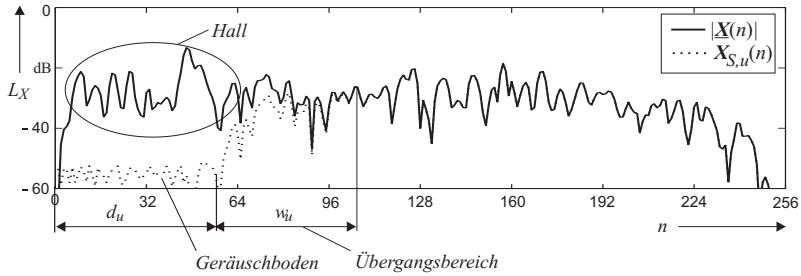

Abbildung 6.2 – Logarithmisch skaliertes Spektrum eines stimmlosen Lautes (stimmloses /s/). Im originalen Spektrum $|\underline{X}(n)|$ erkennt man besonders im unteren Frequenzbereich Störungen durch Hall des Vorgängerlautes; es ist sogar dessen harmonische Struktur zu erkennen. Im synthetischen Spektrum $X_{S,u}(n)$ (punktiert) erkennt man, wie die tieffrequenten Hallstörungen durch den energiearmen Geräuschboden ersetzt werden. Der Übergang erfolgt gleitend.

soll das energiearme Rauschen die Realität eines Lautes nachbilden, bei dem in diesem Frequenzbereich keine Merkmale vorhanden sind. Die Energie des Rauschbodens wird deshalb heuristisch mit $L_{X_{Fl}} = -60$ dB festgelegt. Es ist auch möglich, $L_{X_{Fl}}$ adaptiv einzustellen. Dies wird im Rahmen der Arbeit nicht untersucht. Der Übergang zwischen oberen und unteren Frequenzen verläuft gleitend, d. h., es existieren eine Ein- und eine Ausblendfunktion

$$\begin{aligned} W_{u,L}(f) &= \tfrac{1}{2} + \tfrac{1}{2}\cos\left(2\pi\tfrac{(f-f_{d_u})}{2\cdot f_{w_u}}\right) \quad ; \quad f_{d_u} \le f \le f_{d_u} + f_{w_u} \\ W_{u,H}(f) &= \tfrac{1}{2} + \tfrac{1}{2}\cos\left(2\pi\tfrac{(f-(f_{d_u}+f_{w_u}))}{2\cdot f_{w_u}}\right) \quad ; \quad f_{d_u} \le f \le f_{d_u} + f_{w_u}, \end{aligned} \tag{6.8}$$

die durch ein Von-Hann-Fenster realisiert werden. Die Indizes L und H stehen für engl. low und high und deuten darauf hin, dass das jeweilige Fenster den tief- bzw. den hochfrequenten Bereich ein- bzw. ausblendet. Dabei markieren der Frequenzoffset f_{d_u} den Beginn und f_{w_u} die Breite des Übergangs und legen damit Anfang und Ende der beiden Fensterfunktionen fest. Mit

$$d_u = \mathrm{rd}\left(\frac{f_{d_u}}{\Delta f}\right) \qquad \text{und} \qquad w_u = \mathrm{rd}\left(\frac{f_{w_u}}{\Delta f}\right) \tag{6.9}$$

erhält man die benötigte diskrete Darstellung der Fenster

$$\begin{aligned} W_{u,L}(n) &= \tfrac{1}{2} + \tfrac{1}{2}\cos\left(2\pi\tfrac{(n-d_u)}{2\cdot w_u}\right) \quad ; \quad n = d_u,\ \ldots,\ d_u + w_u \\ W_{u,H}(n) &= \tfrac{1}{2} + \tfrac{1}{2}\cos\left(2\pi\tfrac{(n-(d_u+w_u))}{2\cdot w_u}\right) \quad ; \quad n = d_u,\ \ldots,\ d_u + w_u. \end{aligned} \tag{6.10}$$

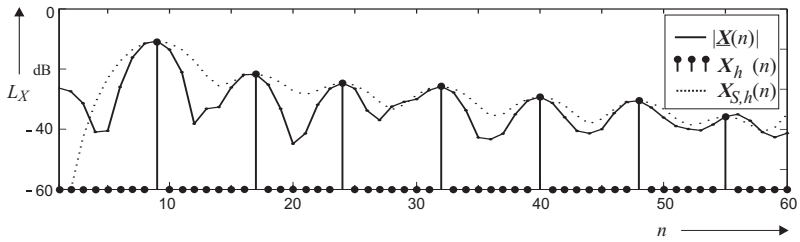

Abbildung 6.3 – Die Darstellung zeigt die ersten 60 (von 256) Spektrallinien eines logarithmisch skalierten Spektrums eines verhallten stimmhaften Lautes (/a/). Im originalen Spektrum $|\underline{X}(n)|$ erkennt man die harmonische Struktur. Zusätzlich ist das Spektrum der Harmonischen $X_h(n)$ (vgl. Gleichung (6.7)) dargestellt, aus dem im unteren Frequenzbereich das synthetische harmonische Spektrum $X_{S,h}$ generiert wird. In diesem Beispiel ist $\gamma = 4$.

Das synthetische Spektrum wird letztlich mit

$$X_{S,u}(n) = \begin{cases} X_{Fl}(n) & ; \quad n = 0, & \ldots,\ d_u - 1 \\ |\underline{X}(n)|\, W_{u,H}(n) + X_{Fl}(n) W_{u,L}(n) & ; \quad n = d_u, & \ldots,\ d_u + w_u - 1 \\ |\underline{X}(n)| & ; \quad n = d_u + w_u, & \ldots,\ \frac{N}{2} - 1 \end{cases} \tag{6.11}$$

berechnet. In Abschnitt 6.3.5 wird beschrieben, wie die optimalen Parameter $f_{d_u} = 500$ Hz und $f_{w_u} = 750$ Hz gefunden werden. Damit wird tieffrequenter Hall unterhalb 1250 Hz mehr und mehr sowie unter 500 Hz vollständig unterdrückt.

6.3.4.2 Stimmhafte Frames

Für Frames, die der VUD als stimmhaft klassifiziert, werden die Ideen **(i)** und **(iii)** aus Abschnitt 6.2 implementiert. Das bedeutet, hochfrequente Anteile werden wieder unverändert belassen (Idee **(iii)**). Tieffrequente Anteile werden aus dem Spektrum von Harmonischen $X_h(n)$ synthetisiert, es entsteht das synthetische harmonische Spektrum $X_{S,h}(n)$ (Idee **(i)**). Der Übergang von tiefen zu hohen Frequenzen wird wieder durch Ein- und Ausblenden der betreffenden Bereiche mit einem Von-Hann-Fenster erreicht. Die notwendigen Parameter d_v und w_v unterscheiden sich allerdings beim stimmhaften Fall von denen des stimmlosen Falls. Das in den tiefen Bereichen generierte synthetische harmonische Spektrum $X_{S,h}$ ist im Wesentlichen eine Interpolation der durch die harmonischen Spektrallinien vorgegebenen Stützstellen. Die Interpolation wird als erster Ansatz wellenförmig gestaltet, angelehnt an die originale spektrale Struktur eines stimmhaften Lautes. Die Wellenform wird durch kleine, sich überlappende Von-Hann-Fenster erreicht. Die einzelnen Fenster erstrecken sich jeweils vom harmonischen

Vorgängerindex ($n_{h,i-1}$) bis zum harmonischen Nachfolgerindex ($n_{h,i+1}$) bei einem bestimmten harmonischen Index $n_{h,i}$

$$W_{n_{h,i}}(n) = \begin{cases} \left(\frac{1}{2} - \frac{1}{2}\cos\left(2\pi \frac{(n-n_{h,i})}{n_{h,i+1}-n_{h,i-1}}\right)\right) & ; \quad n = n_{h,i-1},\ \ldots,\ n_{h,i+1} \\ 0 & ; \quad \text{sonst.} \end{cases} \tag{6.12}$$

Die Synthese erfolgt durch Überlagerung der mit den aktuellen harmonischen Spektrallinien gewichteten Von-Hann-Fenster

$$X_{S,h}(n) = \sum_{i=1}^{i_{\max}} X_h(n_{h,i}) \cdot W^{\gamma}_{n_{h,i}}(n). \tag{6.13}$$

Die Obergrenze $i_{\max}$ ergibt sich wieder aus Gleichung (6.6). Der Exponent γ stellt dabei die Stärke der Welligkeit ein, wobei in Abschnitt 6.3.5 mit $\gamma = 4$ eine günstige Belegung gefunden wird.

Das synthetische stimmhafte Spektrum wird nun ähnlich wie im stimmlosen Fall aus einer Überlagerung der von $X_{S,h}(n)$ in den tiefen Frequenzen und dem originalen Spektrum $|\underline{X}|(n)$ in hohen Frequenzen zusammengesetzt. Der Übergang geschieht, wie erwähnt, analog zu (6.8)

$$\begin{aligned} W_{v,L}(f) &= \tfrac{1}{2} + \tfrac{1}{2}\cos\left(2\pi \tfrac{(f-f_{d_v})}{2\cdot f_{w_v}}\right) && ; \quad f_{d_v} \le f \le f_{d_v} + f_{w_v} \\ W_{v,H}(f) &= \tfrac{1}{2} + \tfrac{1}{2}\cos\left(2\pi \tfrac{(f-(f_{d_v}+f_{w_v}))}{2\cdot f_{w_v}}\right) && ; \quad f_{d_v} \le f \le f_{d_v} + f_{w_v}. \end{aligned} \tag{6.14}$$

Mit den diskreten Frequenzindizes

$$d_v = \mathrm{rd}\left(\frac{f_{d_v}}{\Delta f}\right) \quad \text{und} \quad w_v = \mathrm{rd}\left(\frac{f_{w_v}}{\Delta f}\right) \tag{6.15}$$

erhält man die diskrete Darstellung der Fenster

$$\begin{aligned} W_{v,L}(n) &= \tfrac{1}{2} + \tfrac{1}{2}\cos\left(2\pi \tfrac{(n-d_v)}{2\cdot w_v}\right) && ; \quad n = d_v,\ \ldots,\ d_v + w_v \\ W_{v,H}(n) &= \tfrac{1}{2} + \tfrac{1}{2}\cos\left(2\pi \tfrac{(n-(d_v+w_v))}{2\cdot w_v}\right) && ; \quad n = d_v,\ \ldots,\ d_v + w_v. \end{aligned} \tag{6.16}$$

Das synthetische stimmhafte Spektrum berechnet sich demnach

$$X_{S,v}(n) = \begin{cases} X_{S,h}(n) & ; \quad n = 0, \quad \ldots,\ d_v - 1 \\ |\underline{X}(n)|\, W_{v,H}(n) + X_{S,h}(n) W_{v,L}(n) & ; \quad n = d_v, \quad \ldots,\ d_v + w_v - 1 \\ |\underline{X}(n)| & ; \quad n = d_v + w_v, \quad \ldots,\ \frac{N}{2} - 1. \end{cases} \tag{6.17}$$

In Abschnitt 6.3.5 wird beschrieben, wie die optimalen Parameter $f_{d_v} = 100$ Hz und $f_{w_v} = 3500$ Hz gefunden werden.

6.3.5 Optimales Parameterset

Der zuvor vorgestellte Algorithmus HFA beinhaltet sechs kritische Parameter, deren Einstellung nicht adaptiv erfolgen kann. Diese sind im Einzelnen die Parameter der spektralen Synthese f_{d_u}, f_{w_u}, f_{d_v}, f_{w_v}, γ sowie der VUD-Parameter b_{th}. Es wurden umfangreiche Experimente durchgeführt, die die Zielstellung verfolgten, ein optimales Parameterset für die spektrale Synthese zu finden. Dabei wurde bei Variation eines Parameters die Erkennungsrate für verschiedene Hallbedingungen gemessen. Hier wird nur ein kurzer Überblick über die Ergebnisse gegeben. Einzelheiten sind [Lor08] zu entnehmen. Als Ergebnis wurden die folgenden Einstellungen gefunden:

- $\gamma = 4$: Es wurden die Parameter $\gamma = (1; 2; 4; 8)$ getestet. Um so größer γ ist, um so schmaler werden die Von-Hann-Fenster in (6.12), womit die Welligkeit im synthetischen harmonischen Spektrum $X_{S,h}$ steigt. Die Ergebnisse für $\gamma = (1; 2; 4)$ unterscheiden sich nur leicht mit einem geringen Vorteil für $\gamma = 4$. Bei $\gamma = 8$ wurden für verhallte Testdaten deutliche Einbrüche in der Erkennungsrate festgestellt. Es wird vermutet, dass bei einer Fehlmessung der F_0-Detektion, womit bei verhallten Daten laut Abschnitt 7.4.2 gerechnet werden muss, eine geringere Welligkeit günstiger ist. Die F_0-Fehler wirken sich dadurch nicht so stark aus.
- $f_{d_v} = 100$ Hz und $f_{w_v} = 3500$ Hz : In den Experimenten wurde festgestellt, dass die Variation von f_{d_v} nur geringe Schwankungen in den Ergebnissen hervorruft. Gegenüber anderen Einstellungen entsteht ein leichter Vorteil für $f_{d_v} = 100$ Hz. f_{w_v} erzielte mit 3500 Hz von vornherein gute Ergebnisse. Eine Variation wurde nicht weiter untersucht.
- $f_{d_u} = 500$ Hz und $f_{w_u} = 750$ Hz : Über eine große Anzahl von Experimenten liegt der günstig arbeitende Bereich von f_{d_u} zwischen (100 ... 1000) Hz. $f_{d_u} = 500$ Hz erzielte dabei in den meisten Experimenten die besten Ergebnisse. Für f_{w_u} wurde festgestellt, dass HFA für kleinere Werte als 1000 Hz gute Erkennungsergebnisse generiert, für größere f_{w_u} verschlechtern sich die Erkennungsraten. Es wird vermutet, dass dieser Wert mit den Formantfrequenzen zusammenhängt (vgl. Abbildung 4.1). Eine Erläuterung dazu wird in Abschnitt 6.3.6 gegeben. $f_{w_u} = 750$ Hz wird als optimaler Parameter festgelegt.

Die Einstellung des VUD-Parameters b_{th} erfolgt zunächst heuristisch anhand seines Verhaltens (Abbildung 7.8). Er wird mit $b_{th} = 1$ ($L_{b_{th}} = 0$ dB) festgelegt. In Abschnitt 6.4.3 folgt eine Überprüfung von b_{th} der VUD-Implementierung in HFA durch Spracherkennungsexperimente. Sie bestätigt, dass die heuristische Einstellung tatsächlich ein Optimum darstellt. In Abschnitt 7.6.3 erfolgt eine allgemeine Evaluation mit Messungen der Fehlerrate des VUDs, die ebenfalls feststellt, dass bei diesem Wert gute Ergebnisse erzielt werden.

6.3.6 Behandlung von VUD-Klassifikationsfehlern

Bei den zuvor beschriebenen Algorithmen wird bislang davon ausgegangen, dass der VUD korrekt arbeitet. Wie Abschnitt 7.6.3.2 zu entnehmen ist, arbeitet der hier implementierte VUD selbst bei sauberen Daten mit einer Fehlerrate von etwa 15 %, was auch der Leistungsfähigkeit von anderen in der Literatur gegebenen VUD-Ansätzen entspricht. Bei verhallten Daten werden je nach Stärke der Verhallung Fehlerraten zwischen (25 ... 35) % erreicht. Es darf demnach zunächst geschlussfolgert werden, dass die von HFA implementierten Ideen in den fehlerhaft VUD-klassifizierten Frames ungewünschte Effekte erzeugen. Prinzipiell wird zunächst festgestellt, dass bei Anwendung der zwei unterschiedlichen HFA-Analysemethoden (stimmhaft und stimmlos) auf ein und denselben Frame zwei voneinander verschiedene Merkmalvektoren entstehen. Die oberen Dimensionen besitzen zwar dadurch, dass hochfrequente Abschnitte von HFA unverändert bleiben, die gleichen Werte. In den unteren Dimensionen hingegen weichen die beiden Vektoren erheblich voneinander ab. Sie beschreiben demnach entfernt voneinander liegende Punkte im Merkmalraum. Bildet man Merkmalvektoren einer größeren Anzahl von Frames einer Lautklasse, so formen die beiden Analysearten zwei im Merkmalraum voneinander entfernte Vektorcluster. In einer mehrdimensionalen Häufigkeitsverteilung über eine größere Stichprobe entstehen dadurch zwei Maxima. Aus diesen Gründen ist bei Benutzung von HFA bei einem HMM-Phonemerkenner eine Vektorquantisierung nötig, nach der mindestens zwei Gaußverteilungen eines Gaussian-Mixture-Models gebildet werden. Mit diesen zwei Verteilungen kann die Erkennung von korrekt und fehlerhaft VUD-klassifizierten und im Anschluss entsprechend HFA-analysierten Frames durchgeführt werden. Beim Training der Phonemmodelle bilden sich die beiden Maxima nicht gleichmäßig stark aus. Das fehlerhaft VUD-klassifizierte Cluster bildet sich schwächer aus, da, wie in Abschnitt 7.6.3.2 nachgewiesen wird, davon auszugehen ist, dass der VUD in der Tendenz eher richtig als falsch entscheidet. Die Häufigkeiten werden also im richtigen Cluster größer sein als im falschen. Um das falsche Cluster ebenfalls für die Erkennung brauchbar zu machen, muss die Anzahl der fehlklassifizierten Frames groß genug sein, damit die Verteilungsfunktion relevante statistische Werte erhält. Deshalb ist ein nicht optimal klassifizierender VUD in diesem Zusammenhang sogar nützlich, wenn man davon ausgeht, dass der Idealfall nicht erreichbar ist.

Nach diesen allgemeinen Betrachtungen folgen Gedanken zu den konkret auftretenden Fehlern bei der Erkennung:

Stimmhafter Frame klassifiziert als stimmlos: In diesem Fall werden informationstragende tieffrequente Komponenten von stimmhaften Lauten unterdrückt. Wichtige Informationen gehen dabei verloren. Es verbleiben Informationen beginnend mit dem Übergangsbereich. Dieser wird bei stimmlosen Frames mit $f_{d_u} \ \ldots \ f_{d_u} + f_{w_u} = (500 \ \ldots \ 1250)$ Hz durch die experimentelle Optimierung nach Abschnitt 6.3.5 ein-

gestellt. Laut Abbildung 4.1 geht damit für alle Vokale jeweils der erste Formant verloren. Der zweite (und höhere) hingegen bleibt erhalten. Das bedeutet, auch im fehlerhaft VUD-klassifizierten Cluster sind noch erkennungsrelevante Formantinformationen enthalten. Dies ist vermutlich auch der Grund, warum die Experimente aus Abschnitt 6.3.5 genau diesen optimalen Übergangsbereich erzeugt haben, der mit der Obergrenze 1250 Hz deutlich unterhalb der in HFA-Idee **(iii)** formulierten 2500 Hz liegt.

Stimmloser Frame klassifiziert als stimmhaft: Die stimmhafte spektrale Synthese generiert im unteren Frequenzbereich ein $X_{S,h}(n)$, das auf einem von der F_0-Detektion vorgeschlagenen Wert zwischen $F_{0,\min} \ \ldots \ F_{0,\max}$ beruht. Dabei kann es sich um ein fiktives F_0 oder um eine durch Hall verschliffene harmonische Struktur eines Vorgängervokals handeln (vgl. Abbildung 4.4 (rechts)). Bei verhalltem Training bildet sich dadurch das fehlerhaft VUD-klassifizierte Cluster aus. Es hat allerdings in den unteren Merkmaldimensionen größere Varianzen, da der tieffrequente Hall der vorhergehenden stimmhaften Laute vom jeweiligen Laut abhängige unterschiedliche Charakteristiken besitzt. In diesem Fall ist HFA nicht besser als CFA bei verhalltem Training.

6.4 Evaluation HFA vs. CFA

In diesem Abschnitt wird das Verhalten von HFA mit optimiertem Parameterset (Abschnitt 6.3.5) mit dem Verhalten von CFA verglichen (beide mit MFB-Merkmalen). Eine visuelle Gegenüberstellung der Merkmalmuster beider Techniken bei unverhallten und verhallten Daten wird in Abbildung 6.4 gegeben. In den Darstellungen der Spektrogramme (links) ist sowohl in stimmhaften als auch in stimmlosen Abschnitten die Funktionalität von HFA im Vergleich zu CFA zu beobachten. Niederfrequente stimmhafte Abschnitte werden synthetisiert (erkennbare Wellenform), wodurch Störungen zwischen den Harmonischen unterdrückt werden. Niederfrequente stimmlose Abschnitte werden unterdrückt (erkennbare Lücken im Spektrogramm), wodurch niederfrequenter Hall entfernt wird. Besonders gut ist die (einkalkulierte) Unsicherheit des VUD zu erkennen. Im unverhallten Fall tritt sie vorwiegend in der Anfangs- und Endphase stimmhafter Laute auf, da hier die Energie des Stimmhaften bereits so gering ist, dass sie unter den VUD-Schwellwert sinkt (vgl. Abbildung 7.8). Der VUD klassifiziert folglich die Anfangs- und Endphase stimmhafter Laute als stimmlos. Im halligen Fall sinkt die Energie in der Endphase eines stimmhaften Lautes durch dessen Verschleifung nicht so schnell unter die VUD-Schwelle. Der VUD klassifiziert hier zu spät stimmlos (in Abbildung 6.4 bei $a = 70 \ \ldots \ 80$ bzw. $a = 95 \ \ldots \ 100$ gut zu erkennen). In den Darstellungen der Merkmalmuster (rechts) ist es schwierig, den Effekt der stimmhaften Spektren auszumachen; im Bereich zwischen $a = 110 \ \ldots \ 150$ ist

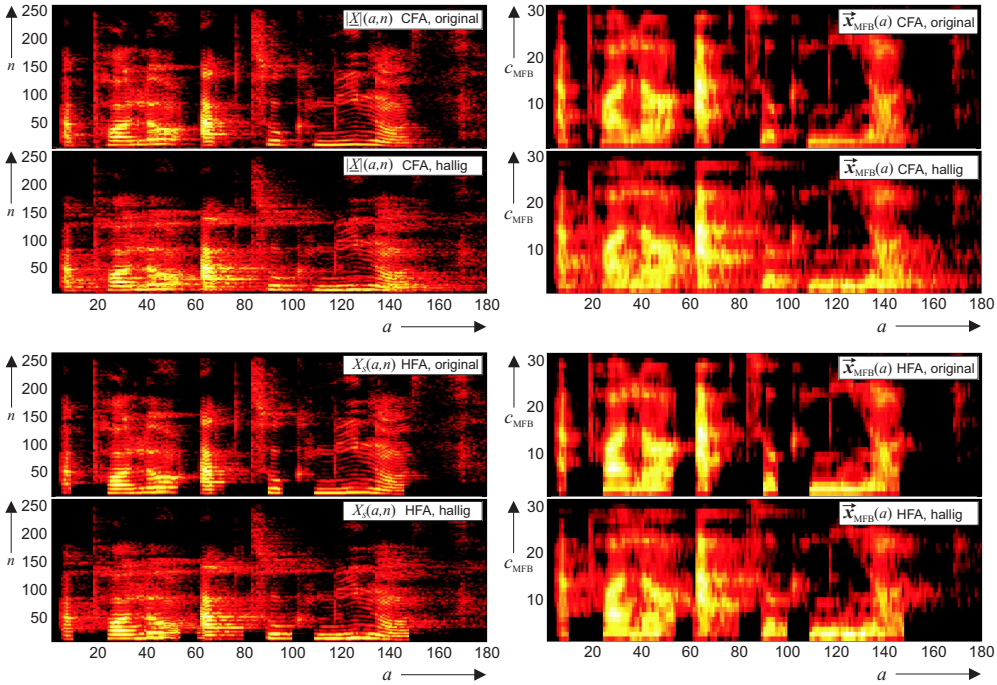

Abbildung 6.4 – Gegenüberstellung von CFA (oben) und HFA (unten) einer Beispieläußerung, die einmal unverhallt und einmal verhallt ($T_{60} = 0,7$ s, SMD = 1 m) analysiert wird. Die Graphiken links zeigen die Spektrogramme (CFA: $|\underline{X}|(a,n)$ und HFA: $X_S(a,n)$). Die Graphiken rechts zeigen die daraus generierten Merkmalmatrizen. Die Unterschiede in den Merkmalmatrizen zwischen original und hallig erscheinen optisch bei HFA geringer als bei CFA. Zur besseren Veranschaulichung ist die Graphik auf der vorderen inneren Umschlagseite farbig abgebildet.

der Vorteil von HFA gegenüber CFA ein wenig auch optisch erkennbar. Die stimmlosen Frames sind bei HFA deutlich zu erkennen; Abschnitte, in denen die Ähnlichkeit zweier stimmloser HFA-Muster (auch optisch) deutlich größer ist als im CFA-Fall, finden sich z. B. bei $a = 100 \ldots 110$ bzw. $a = 150 \ldots 180$. Allgemein gilt, dass die vollständigen Merkmalmuster für HFA einander ähnlicher erscheinen als für CFA, wodurch der zur schlechten Erkennung führende große Unterschied zwischen Trainings- und Testbedingung verringert wird.

Die Arbeitsweise von HFA wird durch Erkennungsexperimente evaluiert. Die Evaluationsbedingungen sind wieder die in Abschnitt 4.5.2 vorgestellte Abhängigkeit der Erkennungsrate von der Nachhallzeit T_{60} sowie die in Abschnitt 4.5.3 vorgestellte Abhängigkeit der Erkennungsrate vom SMD im SMART-Room. Zusätzlich wird das

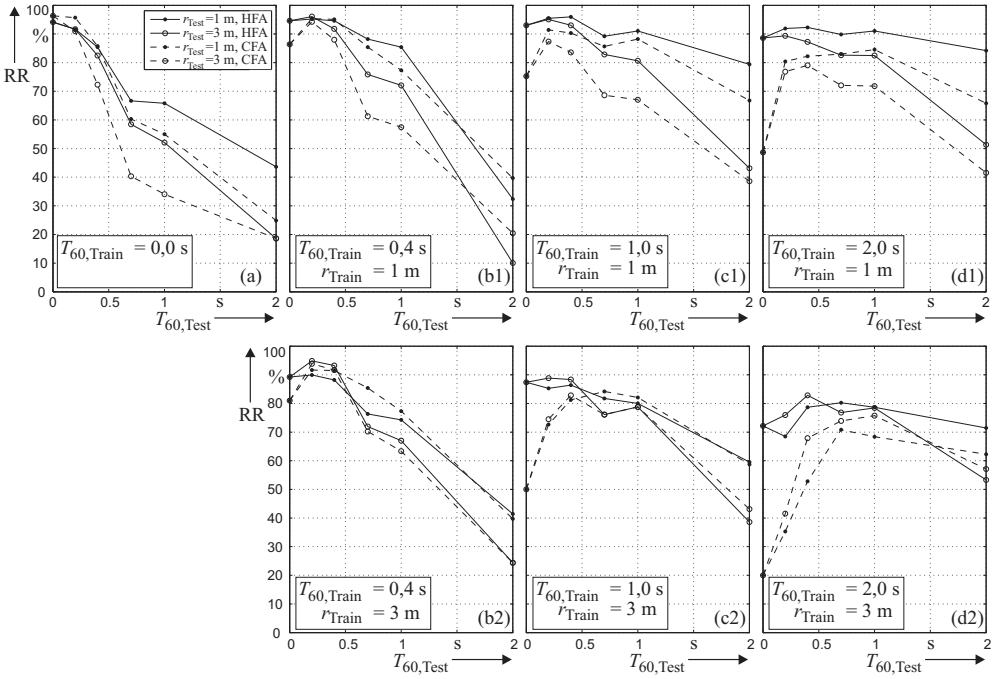

Abbildung 6.5 – Vergleich der Erkennungsergebnisse von HFA und CFA. Die CFA-Ergebnisse sind die selben wie in Abbildung 5.3 (Variante MFB). (a) ungestörte Trainingsdaten. (b1) ... (d1) verhallte Trainingsdaten ($T_{60,\mathrm{Train}} = [400; 1000; 2000]$ ms) mit der SMD-Nahfeldbedingung $r_{\mathrm{Train}} = 1$ m. (b2) ... (d2) verhallte Trainingsdaten mit der SMD-Fernfeldbedingung $r_{\mathrm{Train}} = 3$ m.

Verhalten bei verhalltem Training untersucht.

6.4.1 Abhängigkeit der RR von T_{60}

Die durch HFA erzielten Erkennungsraten für die Abhängigkeit der RR von T_{60} werden mit den durch CFA erzielten Erkennungsraten aus Abbildung 5.3 (Variante MFB) verglichen.

Ungestörte Trainingsdaten (Abbildung 6.5 (a)): Für die ungestörte Trainingsbedingung zeigt sich, dass HFA für keinen bzw. geringen Hall in der Testbedingung zunächst etwas an Erkennungsrate einbüßt. Dieser Effekt ist die Folge des harten Beschneidens der Spektren, was, wie angedeutet, mit dem Verlust von für die Erkennung relevanten Informationen verbunden ist. Die Graphik sagt allerdings auch aus, dass dieser

Verlust für praktische Anforderungen zu verschmerzen ist (< 3 %). Wird die Testbedingung halliger, so erreicht HFA eine signifikante Steigerung der Erkenungsrate (bis zu 20 %). Die in Abbildung 6.4 subjektiv festgestellte größere Ähnlichkeit der Merkmalmuster wird damit nachgewiesen. Obwohl eine signifikante Steigerung erzielt wird, sind die Praxisanforderungen mit Erkennungsraten weit unter 90 % nicht erfüllt (vgl. praktische Einsatzbedingungen in Abschnitt 1.2).

Verhallte Trainingsdaten (Abbildung 6.5 (b1) ... (d2)): Bei verhalltem Training im Nahfeld ($r_{\text{Train}} = 1$ m) steigert sich die Erkennungsrate von HFA (aber auch von CFA) erheblich. Bei CFA fällt auf, dass besonders die zur entsprechenden Trainingsbedingung gehörige Testbedingung gut erkannt wird. Bspw. verschlechtern sich die Ergebnisse bei stärkerer Verhallung der Trainingsbedingung mehr und mehr für die unverhallte Testbedingung. HFA hingegen erzielt nicht nur bessere Werte als CFA, sondern schafft es auch, eine über alle Testbedingungen stabile Steigerung der Erkennungsrate zu erreichen. Als Grund dafür wird insbesondere die Unterdrückung der Tiefen (Hall, vgl. Abbildung 6.2) in den stimmlosen Frames gesehen. Wird verhallt trainiert, entstehen dadurch in diesen Bereichen keine Merkmale (anders als bei CFA). Die Struktur des Merkmalmusters ähnelt damit eher einem unverhallten Fall (vgl. auch Abbildung 6.4), der deshalb auch so gut erkannt wird, obwohl die Testbedingung mit der Trainingsbedingung nicht übereinstimmt.

HFA arbeitet insbesondere gut für den SMD $r_{\text{Test}} = 1$ m. Bei den Trainingsbedingungen $T_{60,\text{Train}} = 1,0$ s bzw. 2,0 s werden durch HFA für das Wohn- und Büroumfeld ($T_{60,\text{Test}} = (0,3 \ \ldots \ 0,8)$ s) bereits die in der Praxis geforderten Erkennungsraten > 90 % erreicht. Aber auch für die Extrembedingung Treppenhaus ($T_{60,\text{Test}} = 2,0$) s ist die erzielte Erkennungsrate von 80 % bzw. 85 % angesichts des Ausgangspunktes von RR = 25 % (Abbildung 6.5 (a) CFA) ein beachtliches Ergebnis. Für die stärker gestörte Fernfeldtestbedingung $r_{\text{Test}} = 3$ m wird die Praxisbedingung RR > 90 % noch nicht vollständig erreicht. Allerdings kommt HFA dieser Bedingung zumindest im Wohn- und Büroumfeld mit RR > 80 % bereits nahe.

Für die Fernfeldtrainingsbedingung $r_{\text{Train}} = 3$ m fallen die Ergebnisse nicht so gut aus. Als Grund dafür wird angenommen, dass die Struktur des Trainingssignals bereits so stark verzerrt ist, dass das Training der Modelle dadurch nicht mehr so diskriminierend ist. Es wird geschlussfolgert, dass die Direktschallkomponente bei verhalltem Training signifikant genug sein muss ($\rightarrow$ Nahfeldbedingung); wie signifikant, beantwortet diese Arbeit nicht vollständig. Allerdings deuten Ergebnisse aus dem folgenden Abschnitt auf eine Antwort hin. Durch diese Ergebnisse motiviert, wird die Fernfeldtrainingsbedingung in allen nachfolgenden Experimenten nicht mehr getestet.

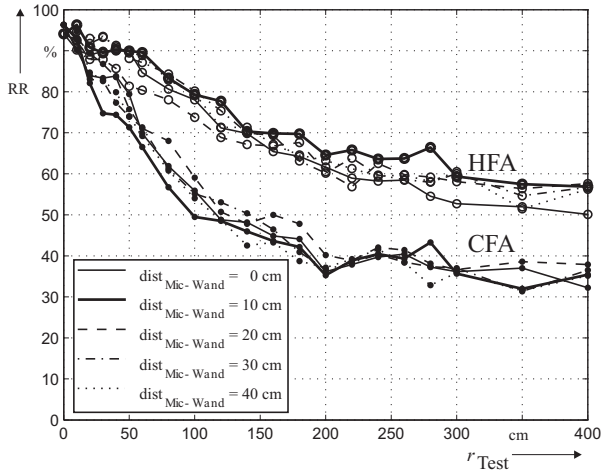

Abbildung 6.6 – Erkennungsrate in Abhängigkeit vom SMD für HFA im Vergleich zu CFA im SMART-Room. Die fünf Graphen der beiden Graphenbündel stehen für verschiedene Abstände des Mikrofons von der Wand ($\text{dist}_{\text{Mic-Wand}} = [0; 10; 20; 30; 40]$ cm). Das Experiment soll zeigen, dass der Einfluss des Wandabstandes des Mikrofons bis auf eine geringe Streuung keinen wesentlichen Einfluss auf die Erkennungsrate hat; sie bleibt in der Tendenz gleich.

6.4.2 Abhängigkeit der RR vom SMD

Analog zum Ausgangspunkt in Abbildung 4.9 (a2) wird auch für HFA die SMD-abhängige Messung im SMART-Room durchgeführt und mit den Ergebnissen für CFA verglichen.

Ungestörte Trainingsdaten (Abbildung 6.6): Bei der ungestörten Trainingsbedingung zeigt sich, dass HFA, wie bei der Abhängigkeit von $T_{60,\text{Test}}$, gegenüber CFA die Erkennungsrate steigert (ebenfalls etwa (20 ... 25) %). Auch hier ist zu bemerken, dass die Steigerung zwar signifikant ist, aber die Praxisanforderung RR > 90 % noch nicht erfüllt wird.

Abbildung 6.6 enthält zusätzlich das sehr interessante Experiment, in dem der Abstand des Mikrofons $\text{dist}_{\text{Mic-Wand}}$ von der Wand variiert wird. Die Sprecherpositionen verändern sich entsprechend, um den SMD beizubehalten. Das Ergebnis zeigt zwar, dass die Erkennungsrate sich von einem $\text{dist}_{\text{Mic-Wand}}$ zum nächsten ändert, allerdings fällt die Änderung nur gering und nicht systematisch aus ($\Delta\text{RR} < 10$ %). Die Unterschiede bei variierenden $\text{dist}_{\text{Mic-Wand}}$ werden nicht mit dem Wandabstand selbst erklärt, sondern vielmehr mit einer unterschiedlichen Gesamtkonstellation der Sprecher-Mikrofon-Strecke. Diese Aussage stützt sich darauf, dass es bis auf wenige Ausnahmen

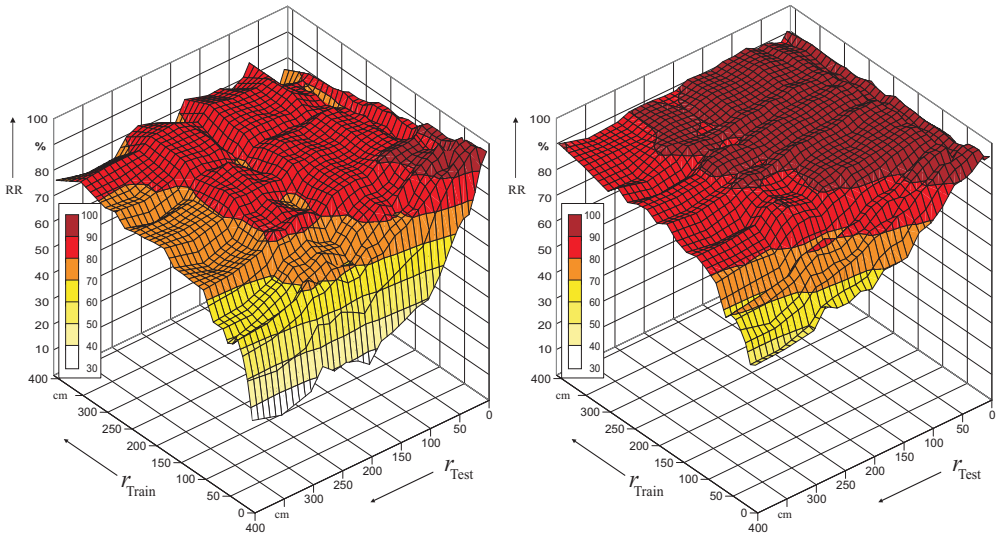

Abbildung 6.7 – Erkennungsrate in Abhängigkeit vom SMD für CFA (links) im Vergleich zu HFA (rechts) im SMART-Room. Der Erkenner wird mit den im SMART-Room aufgenommenen RIRs trainiert. Das dargestellte Netzwerk an der Oberfläche besitzt Abstände von 10 cm, darunter befinden sich auch interpolierte Werte. Trainiert und getestet wurde nur in dem Raster, das zur Verfügung stand (vgl. Bildunterschrift in Abbildung 3.14.).

keine eindeutige Reihenfolge (wechselt) der fünf Graphen eines Graphenbündels in Abbildung 6.6 gibt. Es wird ferner nicht ausgeschlossen, dass die Unterschiede sich durch geringe Messfehler, existierende Störgeräuschbedingungen oder Ungenauigkeiten bei der Einhaltung des SMD bei der RIR-Messung entstehen. Eine konkrete Aussage kann dazu hier nicht getroffen werden. Viel wichtiger ist die Aussage des Experiments, dass die Tendenz der Erkennungsrate unverändert bleibt. Das bestätigt die Richtigkeit der Annahme, dass der SMD und nicht die genaue Position von Sprecher und Mikrofon im Raum das entscheidende Störkriterium ist. Für die folgenden Experimente wird der Abstand $\text{dist}_{\text{Mic-Wand}} = 10$ cm benutzt.

Verhallte Trainingsdaten (Abbildung 6.7): Die Trainingsdaten werden mit den im SMART-Room gemessenen RIRs verhallt und anschließend wird der Erkenner trainiert. Abbildung 6.7 zeigt die Ergebnisse. Diese Graphik beruht auf dem praktischen Hintergrund, dass der Raum bekannt sei und der Erkenner auf diesen trainiert wird. Es ist zu erkennen, dass CFA für die meisten Messpunkte keine ausreichende RR erreicht (obwohl das verhallte Training die Erkennungsrate im Vergleich zum unverhallten Training bereits deutlich steigern kann). HFA hingegen erreicht für eine Vielzahl von

Test- und Trainingsbedingungen Erkennungsraten > 90 %.

Aus den Ergebnissen für HFA lässt sich das Verhalten bezüglich eines optimalen Trainings-SMDs beobachten. In Abschnitt 6.4.1 wird festgestellt, dass das Training mit der Fernfeldbedingung $r_{\text{Train}} = 3$ m ungünstige Ergebnisse liefert, wogegen die Nahfeldbedingung $r_{\text{Train}} = 1$ m erfolgreich arbeitet. In den Ergebnissen für HFA in Abbildung 6.7 ist nun anhand der feineren Unterteilung der Trainings-SMDs zu beobachten, dass die Erkennung für $r_{\text{Train}} = (2 \ldots 3)$ m gut funktioniert (> 90 % für einen Großteil der Testbedingungen). Sowohl bei größeren, aber auch bei kleineren Trainings-SMDs wird das Verhalten der RR wieder ungünstiger. Es wurde bereits festgestellt, dass der SMART-Room mit der einbezogenen Körperabschattung einen Hallradius $r_{\text{R}} \approx (2 \ldots 3)$ m besitzt (vgl. Abschnitt 3.4.3). Die Schlussfolgerung aus diesem Experiment bezüglich des optimalen Trainings-SMDs für HFA lautet demnach: Ein günstiges Verhalten stellt sich bei einem Trainings-SMD ein, der groß genug ist, um der Fernfeldbedingung nahe zu kommen, sich aber noch innerhalb des Hallradius befindet. Will man diese Erkenntnis auf die Ergebnisse aus Abbildung 6.5 übertragen, ist beim SMD immer der Faktor $2 \ldots 3$ einzubeziehen, da diese RIRs ohne Körperabschattung gemessen wurden (angenommene Kugelcharakteristik, $\gamma = 1$). Die Aussagen zur optimalen Traingsdistanz bei HFA können nur als grobe Richtlinie dienen, da die Experimente in dieser Arbeit nicht ausreichend sind, um genauere Schlussfolgerungen zuzulassen.

6.4.3 Einfluss von VUD-Parametern bei HFA

Weitere Untersuchungen beschäftigen sich nachträglich[2] mit der optimalen Einstellung des VUD-Parameters b_{th} bzw. $L_{b_{th}} = 10 \lg b_{th}$ dB. Dazu werden Erkennungsexperimente durchgeführt, die in Abbildung 6.8 im Überblick dargestellt sind. Zielstellung ist es wieder, ein b_{th} zu finden, das für eine große Anzahl von Testbedingungen gut funktioniert. Deshalb werden für eine bestimmte Einstellung von b_{th} die Erkennungsraten wieder bei den Testbedingungen aus Abbildung 6.5 gemessen ($T_{60,\text{Test}} = (0; 200; 400; 700; 1000; 2000)$ ms im Nahfeld $r_{\text{Test}} = 1$ m und im Fernfeld $r_{\text{Test}} = 3$ m). Als Trainingsbedingung wird unverhalltes ((a1) – (a4)) und verhalltes Training ((b1) – (b4)) getestet. Der Erkenner wird jeweils für vier Einstellungen von b_{th} trainiert ($L_{b_{th},\text{Train}} = (-7; -3; 0; 3)$ dB). Eine 3D-Graphik ist nötig, da zur Testphase nicht

[2] Chronologisch erfolgen diese Experimente erst nach denen aus Abschnitt 6.4.1 bzw. 6.4.2, wo, wie in Abschnitt 6.3.5 ausgeführt, die heuristische Einstellung $L_{b_{th}} = 0$ dB benutzt wird. Die im aktuellen Abschnitt vorgestellten Experimente wurden durch Anfragen auf Konferenzen zum Nachweis der günstigen Einstellung von b_{th} motiviert. Thematisch ordnen sie sich jedoch bereits in Abschnitt 6.3.5 ein. Dennoch werden diese Experimente hier aufgeführt, da aufgrund der Chronologie bereits Erfahrungen aus den Experimenten in Abschnitt 6.4.1 bzw. 6.4.2 eingearbeitet werden; so z. B. die Wahl der besonders günstig arbeitenden verhallten Trainingsbedingung $T_{60,\text{Train}} = 1,0$ s, $r_{\text{Train}} = 1$ m, die in Abbildung 6.8 (b1) – (b4) verwendet wird. Die hier vorgestellten Experimente zeigen letztlich, dass die heuristische Festlegung von $L_{b_{th}} = 0$ dB tatsächlich eine gut funktionierende Einstellung ist.

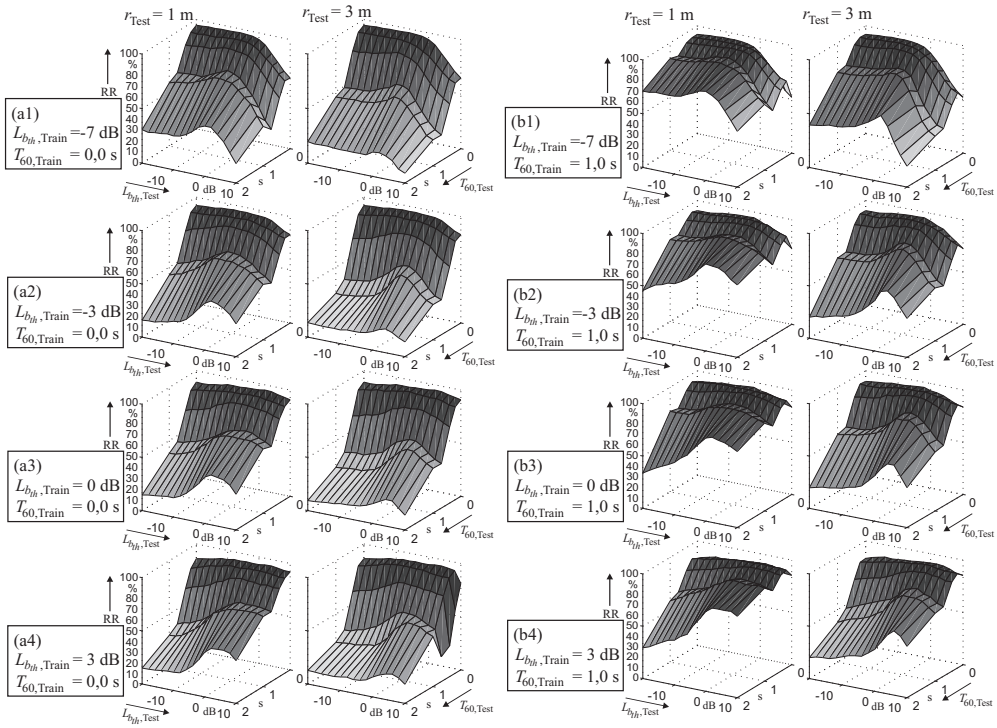

Abbildung 6.8 – Optimierung des VUD-Parameters $L_{b_{th}}$ durch Erkennungsexperimente. (a1) – (a4): unverhallte Trainingsdaten, (b1) – (b4): verhallte Trainingsdaten ($T_{60,\text{Train}} = 1,0$ s, $r_{\text{Train}} = 1$ m). Testbedingungen: Nahfeld $r_{\text{Test}} = 1$ m (links) und Fernfeld $r_{\text{Test}} = 3$ m (rechts), $T_{60,\text{Test}} = (0; 200; 400; 700; 1000; 2000)$ ms. Variation von $L_{b_{th},\text{Test}} = (-20; \ -18; \ \ldots; \ 10)$ dB bei den vier Trainingseinstellungen $L_{b_{th},\text{Train}} = (-7; -3; 0; 3)$ dB (von oben nach unten).

nur das b_{th} eingestellt wird, mit dem vorher auch trainiert wurde, sondern auch weitere $b_{th,\text{Test}}$ getestet werden ($L_{b_{th},\text{Test}} = (-20; \ -18; \ldots; 10)$ dB). Der Grund dieser Variation von $b_{th,\text{Test}}$ ist eine berechtigte Skepsis: Der Schwellwert des VUD basiert auf einer Berechnung der mittleren Energie einer Äußerung über die Zeit nach Gleichung (7.39). In der Realität können die gleichen Äußerungen sehr unterschiedliche mittlere Energien haben, herrührend von unterschiedlicher Sprechgeschwindigkeit, unterschiedlich langen Pausen zwischen Lauten oder Wörtern, unterschiedlicher Prosodie etc. D. h., für zwei Varianten der gleichen Äußerung kann sich die VUD-Schwelle unterschiedlich einstellen. Um dies zu simulieren, wird hier in der Evaluation einfach das b_{th} variiert (zur visuellen Erläuterung vgl. Abbildung 7.8). Abbildung 6.8 sind zunächst zwei Erkenntnisse zu entnehmen: Eine günstige Einstellung ergibt sich bei $L_{b_{th},\text{Train}} = 0$ dB

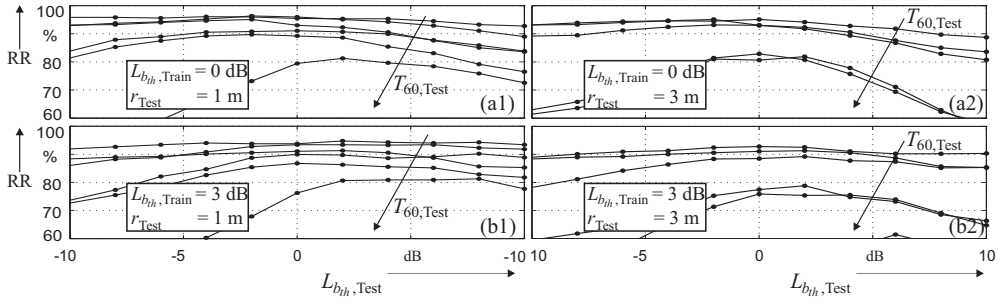

Abbildung 6.9 – Genauerer Auszug aus der Optimierung des VUD-Parameters $L_{b_{th}}$ aus Abbildung 6.8 für die erfolgreiche Einstellung $L_{b_{th},\text{Train}} = (0; 3)$ dB (oben; unten) bei der Trainingsbedingung $T_{60.\text{Train}} = 1,0$ s, $r_{\text{Train}} = 1$ m. Die SMDs der Testbedingungen unterteilen sich in Nahfeld $r_{\text{Test}} = 1$ m (links) und Fernfeld $r_{\text{Test}} = 3$ m (rechts). Die Abhängigkeit der Erkennungsrate von der Nachhallzeit der Testbedingungen $T_{60,\text{Test}} = (0;\ 200;\ 400;\ 700;\ 1000;\ 2000)$ ms ist schematisch als Pfeil eingetragen und soll nur qualitativ auf den Verlauf hinweisen.

bzw. 3 dB; die Ergebnisse für das verhallte Training sind zufriedenstellend. Auf Variation von $b_{th,\text{Test}}$ reagiert das System unkritisch für nicht oder gering verhallte, nicht aber für verhallte Testbedingungen; es ist eine Bevorzugung für $b_{th,\text{Test}}$-Werte, die in der Nähe von $b_{th,\text{Train}}$ liegen, zu beobachten. Die Ergebnisse für die günstigen Einstellungen von $b_{th,\text{Train}}$ (0 bzw. 3 dB) werden in Abbildung 6.9 genauer dargestellt. Man erkennt, dass $L_{b_{th},\text{Train}} = 0$ dB für die meisten Testbedingungen bessere Ergebnisse liefert als $L_{b_{th},\text{Train}} = 3$ dB. Es kann außerdem festgestellt werden, dass bei dieser Einstellung für die meisten Testbedingungen die Erkennungsrate trotz eines schwankenden $L_{b_{th},\text{Test}}$ in weiten Grenzen nahezu konstant bleibt. Mit diesen Einstellungen kann demnach die oben geschilderte Skepsis widerlegt werden. Nach diesen Experimenten wird $L_{b_{th}} = 0$ dB für die Benutzung des VUDs für HFA in dieser Konstellation festgelegt.

6.5 Vergleich mit anderen Ansätzen

Nachdem HFA ausgiebig evaluiert wurde, wird die Methode in diesem Abschnitt mit anderen relevanten Ansätzen verglichen. Nach dem Stand der Technik sind derzeit nur wenige Methoden geeignet, die praxisrelevanten Anforderungen zu erfüllen. Hier werden die drei Varianten TPEFA (Abschnitt 6.5.1.1), Kombination HFA+TPEFA (Abschnitt 6.5.1.2) sowie die IMTF-basierte Enthallung (Abschnitt 6.5.1.3) implementiert und mit HFA (und CFA) verglichen (Abschnitte 6.5.2.1 sowie 6.5.2.2). Zusätzlich wird die Leistungsfähigkeit eines DSBs (Abschnitt 6.5.1.4) mit CFA, HFA sowie der Kombination DSB+HFA (Abschnitt 6.5.1.5) verglichen (Evaluation in Abschnitt 6.5.2.3).

6.5.1 Methoden zum Vergleich

6.5.1.1 TPEFA – TPE Feature Analysis

In Abschnitt 4.2.3 werden die zeitliche Modulation von Sprache in Subbändern, die daraus abgeleiteten TPEs (Temporal Power Envelopes) sowie das Modulationsspektrum $M(\omega_m)$ vorgestellt. Weiterhin wird geschildert, dass für die Spracherkennung wichtige Modulationsfrequenzen im Modulationsspektrum bei $f_m < 20$ Hz, störende bei $f_m > 20$ Hz liegen. In Abschnitt 5.5 werden TPE-basierte Maßnahmen vorgestellt, die ursprünglich Robustheit von ASR unter Geräuschbedingungen erhöhen sollen (z. B. RASTA). Auf dem gleichen Prinzip beruhend wird hier die Tiefpassfilterung von TPEs mit einer Grenzfrequenz $f_{m,g} = 20$ Hz ausgewählt, implementiert und unter Hallbedingungen getestet. Diese Methode wird bereits von Lu et al. [LUA06, LUA07, LUA08] in Verbindung mit einem Spracherkenner (HTK – Hidden Markov Toolkit [HTK02], Erkennung von Sätzen anstelle von Kommandos) sowie halligen Umgebungsbedingungen untersucht. In Abschnitt 4.3.2 wird der tiefpassartige Einfluss des Raumes auf die Modulation von Sprache festgestellt. TPEFA geht davon aus, dass die prinzipielle zeitliche Struktur der TPEs durch den Hall in abgeflachter Form erhalten bleibt und durch den Tiefpass betont wird, der durch Hall eingebrachte hochfrequente Störungen herausfiltert.

Abbildung 6.10 zeigt die Methode als Blockschaltbild. Sie arbeitet zunächst ohne Framing. Das einströmende Signal $x(k)$ wird mit einer Filterbank (CBFB[3] – engl.: Constant-Bandwidth Filterbank) in $C = 64$ Kanäle aufgeteilt. Die Filterbandbreite beträgt, Lu et al. [LUA08] (nach [USFA04]) folgend, 100 Hz. Die Kanäle überlappen sich beim Ein- und Ausblenden der Frequenzen. Am Ausgang des c-ten Kanals entsteht das bandbegrenzte Signal $x_c(k)$, welches mit der vollen Abtastrate f_s zur Verfügung steht. Dabei wird vom Signalmodell in Gleichung (4.6) ausgegangen, das die kanalbezogene Amplitudenmodulation eines Trägers mit der Modulierenden (Einhüllenden) $\mathrm{e}_{s,c}(k)$ für Sprachsignale beschreibt. Da hier das verhallte Signal ansteht, lautet der Index x und nicht s wie in (4.6). Aus $x_c(k)$ soll die Einhüllende $\mathrm{e}^2_{x,c}(k)$ (TPE) des Leistungssignals $x_c^2(k)$ extrahiert werden. Unoki untersucht mehrere Möglichkeiten der Hüllkurvenextraktion (vgl. Abschnitt 5.5.3), wovon hier das Verfahren der Hilberttransformation $\mathcal{H}\{\cdot\}$ benutzt wird. Dabei wird das komplexe analytische Signal $\underline{x}_{a,c}(k)$ gebildet

[3] Bei Lu et al. [LUA08] wird die TPEFA in zwei Filterbankvarianten getestet, die erste benutzt eine Constant-Q-Filterbank (CQFB) und die zweite eine CBFB. Ein konkreter Vergleich mit einer CFA wird nicht angegeben, allerdings werden (Vor-)Experimente erwähnt, in denen eine CFA mit MFCCs unwesentlich schlechter abschneidet als die CQFB, weshalb diese als CFA-Ersatz gewertet wird. Im Vergleich der beiden Filterbankvarianten erzielt die CBFB in [LUA08] zwischen (10 … 18) % bessere Erkennungsraten (für verhallte Daten, im unverhallten Fall unterscheiden sie sich kaum), woraufhin sie für diese Arbeit ausgewählt wird. Diese Ergebnisse korrespondieren in etwa mit den in dieser Arbeit festgestellten Unterschieden in der Erkennungsrate zwischen CFA und TPEFA (vgl. Abbildung 6.15 und 6.16).

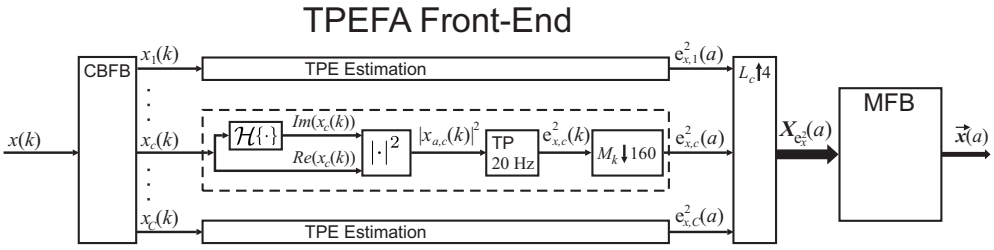

Abbildung 6.10 – Blockschaltbild der Methode TPEFA.

$$\underline{x}_{a,c}(k) = x_c(k) + j\mathcal{H}\left\{x_c(k)\right\}, \tag{6.18}$$

dessen Betrag $|\underline{x}_{a,c}(k)|$ der Einhüllenden von $x_c(k)$ bzw. dessen Betragsquadrat $|\underline{x}_{a,c}(k)|^2$ der Einhüllenden von $x_c^2(k)$ entspricht. Da diese Einhüllenden noch hochfrequente Schwankungen aufweisen, werden sie mit einem Tiefpass ($\mathrm{TP}_{20\ \mathrm{Hz}}$) gefiltert, um letztlich den TPE zu erhalten

$$\mathrm{e}^2_{x,c}(k) = \mathrm{TP}_{20\ \mathrm{Hz}}\left(\left|\underline{x}_{a,c}(k)\right|^2\right). \tag{6.19}$$

Die Grenzfrequenz des $\mathrm{TP}_{20\ \mathrm{Hz}}$ von 20 Hz ist, wie oben erwähnt, durch die Aussagen in Abschnitt 4.2.3 motiviert. Um $\mathrm{e}^2_{x,c}(k)$ von f_s auf die für die Spracherkennung benötigte Framerate (bzw. $FI = 160$, vgl. Gleichung (2.4)) zu dezimieren, wird mit $M_k = 160$ unterabgetastet, sodass $\mathrm{e}^2_{x,c}(a)$ entsteht. Dies ist ohne Weiteres möglich, da der vorgeschaltete $\mathrm{TP}_{20\ \mathrm{Hz}}$ die Verletzung der Abtastbedingung verhindert:

$$20\ \mathrm{Hz} \;<\; \frac{1}{2}\frac{f_s}{M_k} \;=\; 50\ \mathrm{Hz}. \tag{6.20}$$

Somit besitzt der $\mathrm{TP}_{20\ \mathrm{Hz}}$ in dieser Konstellation drei Funktionen: er glättet die Einhüllende, er eliminiert störende hohe Modulationsfrequenzen und er sorgt als Anti-Aliasing-Filter für die Einhaltung der Abtastbedingung. Um die Methode mit CFA bzw. HFA zu vergleichen, muss sichergestellt werden, dass die Effekte in der Erkennungsrate nur von der Benutzung der TPEs stammen. Deshalb wird die gleiche MFB wie bei CFA benutzt, die einen Eingang von 256 Spektrallinien besitzt. Um aus den 64 TPEs ein passendes spekrales Signal $X_{\mathrm{e}^2_x}(a,n)$ zu generieren, wird die Frequenzabtastung um den Faktor $L_c = 4$ erhöht, was einer (Frequenz-)Überabtastung entspricht. Die Interpolation mit den drei Werten zwischen zwei kanalbezogenen Spektrallinien $\mathrm{e}^2_{x,c}(a)$ und $\mathrm{e}^2_{x,c+1}(a)$ erfolgt linear. Die Benutzung des $\mathrm{TP}_{20\ \mathrm{Hz}}$, die zeitliche Unterabtastung von $\mathrm{e}^2_{x,c}(k)$ sowie die spektrale Interpolation für die MFB-Schnittstelle sind leichte Abweichungen der Implementierung in [LUA08], die auf die Funktionsweise keinen Einfluss haben.

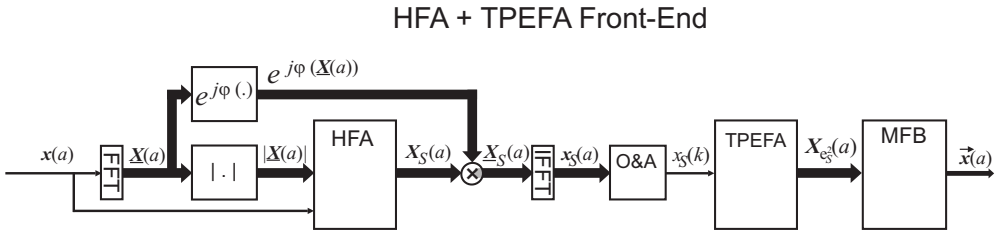

Abbildung 6.11 – Blockschaltbild der Kombination HFA + TPEFA.

6.5.1.2 Kombination HFA + TPEFA

Sowohl HFA als auch TPEFA haben Vor- und Nachteile. Um die Vorteile beider Methoden zu nutzen, wird die Kombination von HFA+TPEFA untersucht.

Ein Blockschaltbild von HFA+TPEFA ist in Abbildung 6.11 dargestellt. Zunächst wird das Signal mit HFA analysiert (ohne MFB), es entstehen die synthetischen Spektren $X_S(a,n)$. Um TPEFA anschließend zu benutzen, muss $X_S(a,n)$ in ein Zeitsignal umgewandelt werden. Dazu werden die einzelnen synthetischen Spektren mit der IFFT in eine Folge von Zeitsignalframes $x_S(a,k)$ transformiert. Wichtig ist dabei noch, dass vor der IFFT aus Betrag ($X_S(a,n)$ ist ein Betragsspektrum) und Phase φ ein komplexes Spektrum $\underline{X}_S(a,n)$ generiert wird. Die Phase wird aus $\underline{X}(a,n)$ gewonnen. Die Zeitsignalframes $x_S(a,k)$ werden mit einem Overlap-and-Add-Algorithmus (O&A) [OS04] zu einem kontinuierlichen Signal $x_S(k)$ zusammengesetzt. Aus diesem Grund muss das Frameintervall von $FI = 160$ auf $FI = 256$ (halbe Framelänge, vgl. Gleichung (2.4)) angepasst werden. Im Anschluss erfolgt TPEFA sowie die MFB, um die Ergebnisse wieder vergleichbar zu machen. Eine umgekehrte Reihenfolge, d. h. zuerst TPEFA und danach HFA, ist nicht möglich, da TPEs weder resynthetisiert noch aus TPEs Harmonische zurückgewonnen werden können.

6.5.1.3 IMTF-basierte Enthallung

Zum Vergleich soll ein Enthallungsverfahren getestet werden. Die Methode der Enthallung der TPEs mit der inversen MTF (IMTF) nach Unoki (vgl. Abschnitt 5.5.3) ist derzeit die einzige, die eine blinde Enthallung so durchführt, dass sie praktischen Implementierungsanforderungen entspricht. Dies sind:

- **Rechenleistung** Die IMTF-Methode benötigt im Vergleich zu anderen Methoden der blinden Enthallung (z. B. Nakatanis Systeminversion) nur geringe Rechenleistung. Insgesamt kann die Rechenleistung als moderat eingestuft werden. Sie verteilt sich auf die drei Blöcke TPE-Extraktion (Rechenleistung wie TPEFA), Schätzung von T_{60} (moderate Rechenleistung) sowie die eigentliche IMTF-Filterung (sehr geringe Rechenleistung, siehe unten).

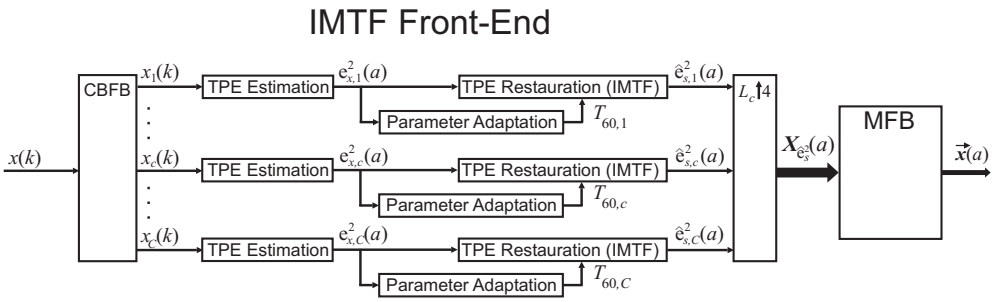

Abbildung 6.12 – Blockschaltbild der IMTF-basierten Methode.

- **Speicheranforderungen** Die Speicheranforderungen sind größer als für CFA, allerdings im Vergleich zu anderen Enthallungsmethoden ebenfalls gering.
- **Positionsunabhängigkeit, Stabilität** Die Methode verhält sich unkritisch zu Änderungen der RIR, wie bspw. der Änderung der Sprecherposition, solange der SMD der Fernfeldbedingung entspricht und T_{60} konstant ist (wovon bei gleichem Raum in etwa ausgegangen werden kann). Dies ist ein Vorteil gegenüber Methoden der Systeminversion, die extrem kritisch auf geringe Änderungen der RIR reagieren (ein Sprecher bzw. weitere Objekte dürften sich während des Sprechens im Raum nicht bewegen).
- **Adaptionszeit** Die Adaption bezieht sich nur auf die Bestimmung von T_{60}. Dabei ist es ausreichend, eine kurze Äußerung (z. B. Kommandowort/ -phrase) zu erfassen (vgl. Abschnitt 5.5.3, Schätzung von T_{60} von Unoki). Außerdem kann z. B. bei einem Haushaltsgerät davon ausgegangen werden, dass es in einem Raum steht, dessen T_{60} sich nur in geringen Grenzen ändert. D. h., T_{60} kann über einen langen Zeitraum und über sämtliche akustischen Ereignisse gemittelt werden, wodurch ein genauer Messwert entsteht.
- **Echtzeitfähigkeit** Durch die kurze Adaptionszeit sowie die moderate Rechenleistung der Methode kann sie sowohl in Echtzeit reagieren als auch in einem Echtzeitsystem implementiert werden. Besonders günstig ist die Implementieung des FIR-Filters zweiter Ordnung (siehe unten).

Die Tatsache, dass die IMTF-Methode nur im Fernfeld arbeitet, ist ein Nachteil gegenüber anderen Methoden, die z. B. mit der inversen RIR arbeiten. Ein weiterer Nachteil ist, dass die positiven Effekte von verhalltem Training nicht genutzt werden können, da dies bei der IMTF-Methode nicht sinnvoll ist.

In einem DSP-System würde man Gleichung (5.12) bzw. (5.13) mit der FFT lösen. Aufgrund der Einfachheit von $\mathrm{e}_h^2(t)$ im Fernfeld ergibt sich hier aber eine wesentlich günstigere Implementierung im Zeitbereich. Diese geht von Gleichung (4.31) aus,

notiert diese aber in diskreter Schreibweise

$$\begin{aligned} \mathrm{e}_h^2(k) &= a_h \tfrac{1}{r_\mathrm{R}^2} \tfrac{13{,}8}{T_{60}} \cdot e^{-\frac{13{,}8}{T_{60}} k \cdot \Delta t} = \beta \cdot e^{-\frac{13{,}8}{T_{60}} \frac{k}{f_s}} \\ &= \beta \cdot \alpha^k \end{aligned} \tag{6.21}$$

mit

$$\beta = a_h \frac{1}{r_\mathrm{R}^2} \frac{13{,}8}{T_{60}} \quad , \quad \alpha = e^{-\frac{13{,}8}{T_{60}} \frac{1}{f_s}} \quad , \quad k \geq 0. \tag{6.22}$$

Gleichung (6.21) entspricht einem IIR-Filter erster Ordnung. Dessen inverses System wird über die $\mathcal{Z}$-Transformation berechnet. Diese ergibt mit üblichen Korrespondenzen [OS04] die MTF (nicht normiert, deshalb die Notation $\underline{M}_h$ anstelle von $\underline{m}$, vgl. Gleichung (4.27))

$$\begin{aligned} \underline{M}_h(z) &= \mathcal{Z}\left\{\mathrm{e}_h^2(k)\right\} = \mathcal{Z}\left\{\beta \cdot \alpha^k\right\} \\ &= \beta \cdot \tfrac{1}{1-\alpha z^{-1}}. \end{aligned} \tag{6.23}$$

Daraus kann leicht die inverse MTF berechnet werden

$$\underline{M}_h^{-1}(z) = \frac{1}{\beta}(1 - \alpha z^{-1}). \tag{6.24}$$

Aus Gleichung (5.12) wird damit

$$\underline{\hat{M}}_s(z) = \underline{M}_x(z) \cdot \frac{1}{\beta}(1 - \alpha z^{-1}). \tag{6.25}$$

Überführt man (5.12) mit der inversen $\mathcal{Z}$-Transformation in den Zeitbereich, ergibt sich

$$\hat{\mathrm{e}}_s^2(k) = \mathrm{e}_x^2(k) * \mathrm{e}_{\mathrm{IMTF}}^2(k) \tag{6.26}$$

mit

$$\mathrm{e}_{\mathrm{IMTF}}^2(k) = \mathcal{Z}^{-1}\left\{\frac{1}{\beta}(1 - \alpha \cdot z^{-1})\right\} \tag{6.27}$$

$$= \frac{1}{\beta} \cdot (\delta(k) - \alpha \cdot \delta(k-1)) \tag{6.28}$$

$$= \frac{1}{a_h} r_\mathrm{R}^2 \frac{T_{60}}{13{,}8} \left(\delta(k) - e^{-\frac{13{,}8}{T_{60}} \frac{1}{f_s}} \cdot \delta(k-1)\right). \tag{6.29}$$

Das bedeutet, dass die TPE-Enthallung mit der IMTF durch ein FIR-Filter erster Ordnung realisiert werden kann, was eine sehr einfache Implementierung darstellt. Sie nimmt eine nicht erwähnenswerte Rechenleistung in Anspruch und kann in Echtzeit berechnet werden. Als Eingangsparameter wird nach wie vor T_{60} benötigt.

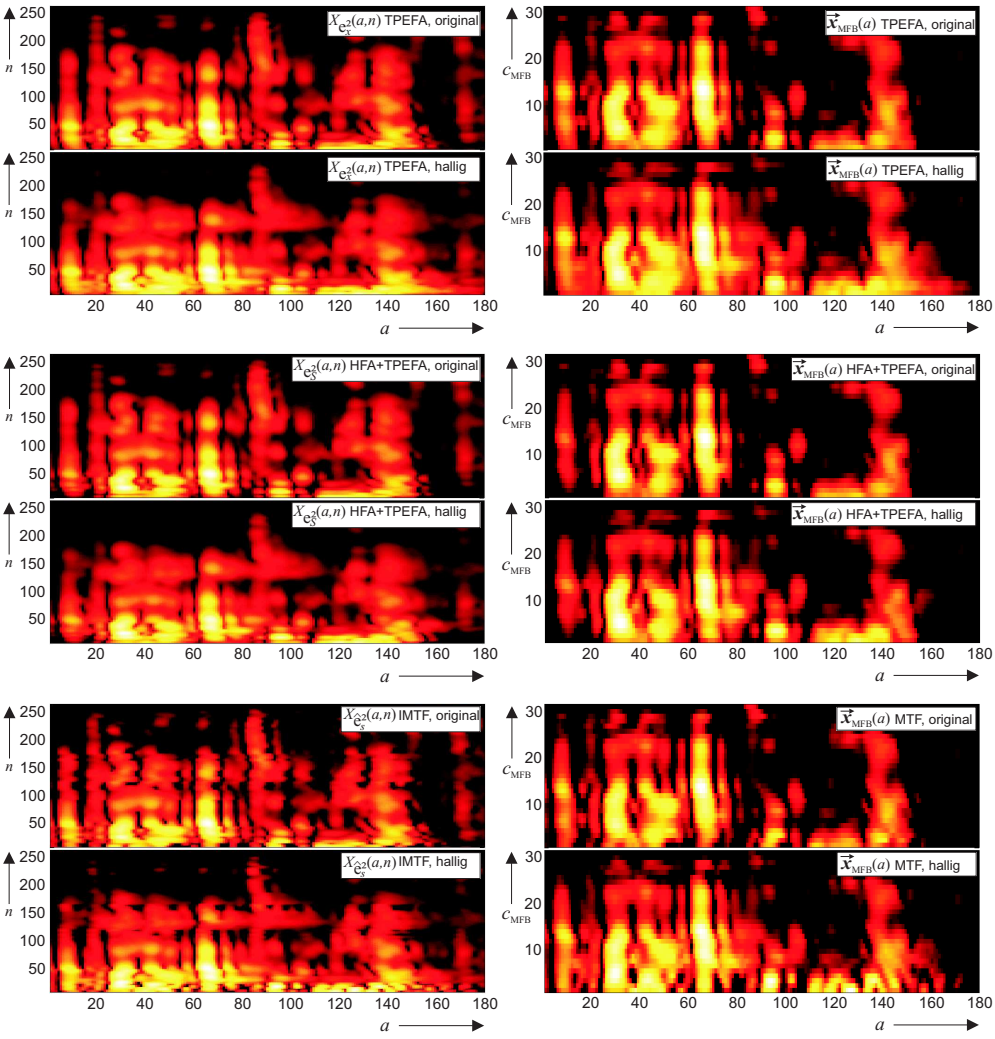

Abbildung 6.13 – Gegenüberstellung von TPEFA (oben), HFA+TPEFA (mitte) und IMTF-Methode (unten) für die Beispieläußerung aus Abbildung 6.4 nach dem gleichen Schema: einmal unverhallt und einmal verhallt ($T_{60} = 0,7$ s, SMD = 1 m). Links: Spektrogramme der TPEs (TPEFA: $X_{e_x^2}(a,n)$, HFA+TPEFA: $X_{e_S^2}(a,n)$ und IMTF: $X_{\hat{e}_S^2}(a,n)$). Rechts: daraus generierte Merkmalmatrizen. Zur besseren Veranschaulichung ist die Graphik auf der hinteren inneren Umschlagseite farbig abgebildet.

6.5.1.4 DSB – Delay-and-Sum Beamformer

Um die Annahme zu überprüfen, dass DSBs für die Spracherkennung in halligen Umgebungen nur eine begrenzte Wirkung haben (vgl. Abschnitt 5.3.3), wird ein DSB implementiert und seine Wirkung mit HFA verglichen. Als DSB-Implementierungen wurden die drei Geometrien in Abbildung 6.14 (a) bis (c) in Erwägung gezogen. Abbildung 6.14 (d) beschreibt die damit erreichten Bündelungsmaße. Man erkennt deutlich, dass DSBs keine effektive Richtwirkung für Wellenlängen erreichen, die groß im Vergleich zu den Distanzen zwischen den Mikrofonen sind (betrifft tiefe Frequenzen). Erhöht man die Anzahl der Mikrofone, ohne die Gesamtausmaße des Arrays zu verändern, wird die Richtwirkung für tiefe Frequenzen sogar noch verringert, obwohl dadurch der Maximalwert des Bündelungsgrades erhöht wird (vgl. Abbildung 6.14 (d) bzw. [BW01]). Um bereits eine Richtcharakteristik für besonders tiefe Frequenzen zu erreichen, können nur Arrays mit größeren Ausmaßen benutzt werden. Jedoch sind besonders große Ausmaße aus praktischen Gesichtspunkten, wie dem Einbau in ein sprachgesteuertes Gerät, unerwünscht. Für die Experimente wurde die Geometrie aus Abbildung 6.14 (c) benutzt, die einen Kompromiss aus Praktikabilität und Leistungsfähigkeit beschreibt. Sie stellt ein Linien-Array dar, dass keine äquidistanten Abstände besitzt, um besonders große Nebenkeulen zu unterdrücken. Dieser Effekt wurde im Rahmen der Diplomarbeit [Feh06] optimiert.

Wendet man die Faltung in Gleichung (3.65) auf den DSB in Gleichung (5.1) an, entsteht

$$x_{\mathrm{DSB}}(k) = \sum_{m=1}^{M} (s * h_m)(k - \kappa_\tau), \tag{6.30}$$

wobei h_m die RIR zwischen Sprecher und m-ten Mikrofon entspricht. Mit der Annahme, dass die $h_m(k)$ lineare Systeme sind, kann Gleichung (6.30) auch wie folgt

$$\begin{aligned} x_{\mathrm{DSB}}(k) &= (s * \underbrace{\sum_{m=1}^{M} h_m(l - \kappa_\tau)}_{h_{\mathrm{DSB}}})(k) && (6.31)\\ &= (s * h_{\mathrm{DSB}})(k) && (6.32) \end{aligned}$$

geschrieben werden. Das bedeutet, dass zum Simulieren einer DSB-Funktionalität nicht die M RIRs wie in (6.30) mit den Sprachsignalen gefaltet werden müssen, sondern dass auch eine einzelne RIR h_{DSB} wie in Gleichung (6.32) benutzt werden kann, die aus der Summe der $h_m(k)$ berechnet wird. In den hier durchgeführten Experimenten kann $\kappa_\tau = 0$ gesetzt werden, da nur Messungen in Hauptrichtung durchgeführt wurden.

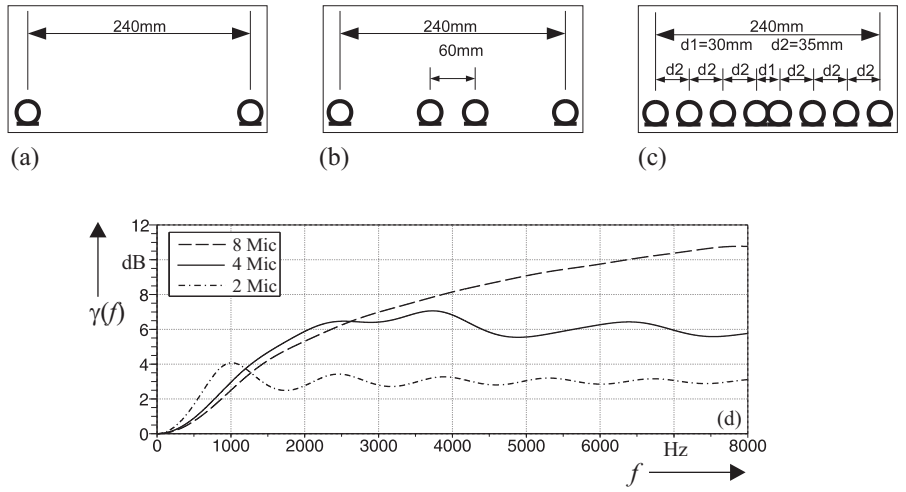

Abbildung 6.14 – Delay-and-Sum Beamformer. Teilabbildungen (a) bis (c) beschreiben graphisch die geometrische Struktur von drei Mikrofonarrays (2, 4 und 8 Kanäle). Teilabbildung (d) zeigt die simulierten Bündelungsmaße $\gamma(f)$ nach Gleichung (3.61) der drei Arrays bei Verschaltung als DSB und unter der Annahme, dass es sich um Einzelmikrofone mit Kugelcharakteristik handelt. (Graphik erstellt von Thomas Fehér).

6.5.1.5 Kombination DSB + HFA

Die Kombination von DSB und HFA testet, ob das Zusammenwirken beider Methoden eine Addition der Vorteile beider Methoden verursacht. Technisch wird die Kombination einfach durch die Reihenschaltung von DSB und HFA erreicht. Dies ist ohne Weiteres möglich, da das Ausgangssignal des DSBs $x_{\mathrm{DSB}}(k)$ ein Zeitbereichssignal ist.

6.5.2 Experimente

Die Erkennungsexperimente zum Vergleich von HFA mit anderen Methoden unterteilen sich in zwei Gruppen. Zuerst werden die Methoden CFA, HFA, TPEFA, HFA+-TPEFA sowie IMTF bei unverhalltem und verhalltem Training getestet und verglichen, wobei in die Abhängigkeiten von T_{60} (Abschnitt 6.5.2.1) und vom SMD (Abschnitt 6.5.2.2) unterschieden wird. Die zweite Gruppe bildet einen Vergleich von CFA, DSB, HFA und HFA+TPEFA (Abschnitt 6.5.2.3). Dabei war es aufgrund des DSBs nötig, neue raumakustische Umgebungsbedingungen mehrkanalig zu messen, da zuvor nur einkanalig gearbeitet wurde. Die Ergebnisse sind somit nur tendenziell, aber nicht direkt mit den Ergebnissen aus der ersten Gruppe zu vergleichen. Es wird versucht, die Abhängigkeit von T_{60} und vom SMD zusammenhängend darzustellen.

In den Darstellungen der Spektrogramme und der Merkmalmatrizen in Abbildung 6.13 wird die Funktionsweise der drei Front-End-Methoden TPEFA, TPEFA+HFA und IMTF visualisiert. Die Glättung durch den $\mathrm{TP}_{20\ \mathrm{Hz}}$ ist im Vergleich zur CFA bzw. HFA (Abbildung 6.4) deutlich erkennbar.

Vergleicht man TPEFA und HFA+TPEFA, so ist der zusätzliche Einfluss von HFA in den unteren Frequenzen zu beobachten. Er eliminiert wieder die tiefen Frequenzen in stimmlosen Abschnitten und führt letztlich dazu, dass die beiden Merkmalmuster für HFA+TPEFA optisch ähnlicher erscheinen als die für TPEFA (analog HFA vs. CFA in Abbildung 6.4). Besonders gut zu beobachten ist dies in den Abschnitten $a = 0 \ldots 70$ bzw. $a = 110 \ldots 180$. Für den stimmhaften Fall ist hier im Gegensatz zu Abbildung 6.4 ebenfalls ein ähnlicheres Muster durch den Einfluss von HFA sichtbar; gut zu beobachten zwischen $a = 25 \ldots 60$.

Da es sich in der Abbildung nicht um eine Fernfeldbedingung handelt (SMD beträgt 1 m), wird erwartet, dass die IMTF-Methode nicht optimal arbeitet. Die Methode schätzt eine Nachhallzeit, auf deren Grundlage eine Enthallung durchgeführt wird, die auf der Fernfeldannahme beruht. Deshalb wird erwartet, dass die IMTF-Methode im Nahfeld die Struktur der Merkmale eher stört als einen positiven Effekt erzeugt. Man erkennt im Spektrum $X_{\hat{\mathrm{e}}_s^2}(a, n)$ zeilenweise Sprünge von einem n zum nächsten. Dieser Effekt beruht auf Ungenauigkeiten der Schätzung von T_{60}, die für jeden Kanal einzeln erfolgt und damit jeweils unterschiedlich ausfallen kann. Es entstehen die Effekte Optimal-, Über- und Unterrestauration, wie sie in Gleichung (5.16) beschrieben werden. Eine zusätzliche Fehlerquelle ist die Tatsache, dass auch die T_{60}-Schätzung auf der Fernfeldbedingung beruht und demnach im Nahfeld ebenfalls kritisch ist. Im Fernfeld arbeitet sie gut für große T_{60}, nicht aber für kleine[4], was durch die größere Anzahl an Sprüngen in $X_{\hat{\mathrm{e}}_s^2}(a, n)$ für das originale Signal ($T_{60} = 0$ s) zum Ausdruck kommt. Diese Sprünge sind durch die cepstrale Glättung in der MFB im Merkmalmuster (rechts) nicht mehr vorhanden. Eventuell wirken sie deshalb im Einzelfall nicht störend.

Um die drei Methoden zu testen, steht wieder das Szenario aus Abschnitt 6.4 zur Verfügung.

6.5.2.1 Abhängigkeit der RR von T_{60}

Die Ergebnisse dieser Experimente werden in Abbildung 6.15 zusammengefasst. Die Erkennungsrate wird bei den beiden Testbedingungen Nahfeld ($r_{\mathrm{Test}} = 1$, oben) und Fernfeld ($r_{\mathrm{Test}} = 3$ m, unten) gemessen. Das Training wird sowohl mit sauberen als auch mit verhallten Sprachdaten durchgeführt, wobei nach den Erkenntnissen zu verhalltem Training mit der Fernfeldbedingung $r_{\mathrm{Train}} = 3$ m (Abbildung 6.5, unten) hier

[4] Quelle: Persönliche Korrespondenz mit Lu und Unoki. Unokis Veröffentlichungen [UFSA04, USFA04] bestätigen diese Aussage zunächst nicht.

nur noch die Nahfeldtrainingsbedingung $r_{\text{Train}} = 1$ m untersucht wird. Zum Vergleich werden die Ergebnisse für CFA und HFA mit abgebildet.

Ungestörte Trainingsdaten (Abbildung 6.15 (a1) und (a2))

TPEFA: Sowohl im Vergleich zu CFA als auch zu HFA erzielt TPEFA teilweise bedeutende Steigerungen der Erkennungsrate. Dies wird bereits im unverhallten Testfall deutlich, wo eine Steigerung von 96 % auf 99 % (CFA vs. TPEFA) erzielt wird[5]. Die geforderte Praxisbedingung RR = 90 % für variierende Wohn- und Büroumgebungen wird dennoch nicht erreicht. Trotzdem ist durch das Konzept des Tiefpasses $\text{TP}_{20\text{ Hz}}$ für alle Testbedingungen ein bemerkenswerter Vorteil entstanden. Im Vergleich zu den unverhallten sind die verhallten TPEs in ihrer Feinstruktur beschädigt. Die Grobstruktur, die mit linguistischen Ereignissen korreliert, bleibt jedoch, wie schon erwähnt, erhalten, wenn sie auch durch das Tiefpassverhalten der MTF abgeflacht ist (Abbildung 4.5). Der Tiefpass $\text{TP}_{20\text{ Hz}}$ eliminiert die Feinstruktur und extrahiert die Grobstruktur aus den TPEs, aus der sowohl für das Training als auch für die Erkennung die wichtigsten Informationen gewonnen werden.

HFA+TPEFA: In Abschnitt 6.4 wurde festgestellt, dass HFA einige nützliche Informationen eliminiert, wodurch die Erkennungsrate für wenig hallige Signale zunächst geringfügig sinkt. Dieser Effekt ist auch hier gegenüber TPEFA zu beobachten. Ab einer signifikanten Halligkeit ist die Eliminierung der Störung der betreffenden Bereiche jedoch wichtiger als die verlorenen Informationen, wodurch die Erkennungsrate gegenüber TPEFA gesteigert wird. Besonders gut ist dies für den größeren $r_{\text{Test}} = 3$ m in Abbildung (a2) zu beobachten. Das Signal ist durch den größeren SMD noch halliger. Das Hinzuziehen von HFA zu TPEFA hat einen ähnlichen Effekt wie bei der Implementierung von HFA zu CFA. Die Erkennungsrate ist insgesamt höher als für HFA, bestimmt durch den Vorteil, den TPEFA schafft.

IMTF: Die IMTF-Methode versucht, die TPEs zu enthallen; sie soll also TPEFA für verhallte Daten verbessern. Die Erwartung, dass die IMTF-Methode im Nahfeld die Ergebnisse eher verschlechtert als verbessert, wird durch dieses Experiment (a1) nicht nachgewiesen. Sie arbeitet nicht schlechter als TPEFA für hallige Umgebungen. Allerdings ist $r_{\text{Test}} = 1$ m auch nicht sehr weit vom Fernfeld entfernt[6]. Bei geringer Halligkeit ergibt sich jedoch eine leichte Verschlechterung (ca. 5 %), was, wie erwähnt, auf die ungenauere Schätzung von T_{60} in diesem Bereich zurückgeführt wird. Versucht die Methode mit einem geschätzten T_{60}, ungestörte Daten zu enthallen, werden Störungen eingebracht, die sich in der verminderten Erkennungsrate auswirken. Für $r_{\text{Test}} = 3$ m in (a2) wird durch die IMTF-Methode eine Steigerung für hallige Umge-

[5] Im Bereich über 95 % ist eine Steigerung um 3 % eine erhebliche Verbesserung. Rechnet man in Fehlerraten, verbessert sich bei diesem Beispiel das Ergebnis um den Faktor 4 von ER = 4 % auf ER = 1 %.

[6] Typische Hallradien liegen z. B. mit $\gamma > 4$ bei $(1 \ldots 2)$ m, vgl. Verteilung der Hallradien in Abschnitt 3.4.2.

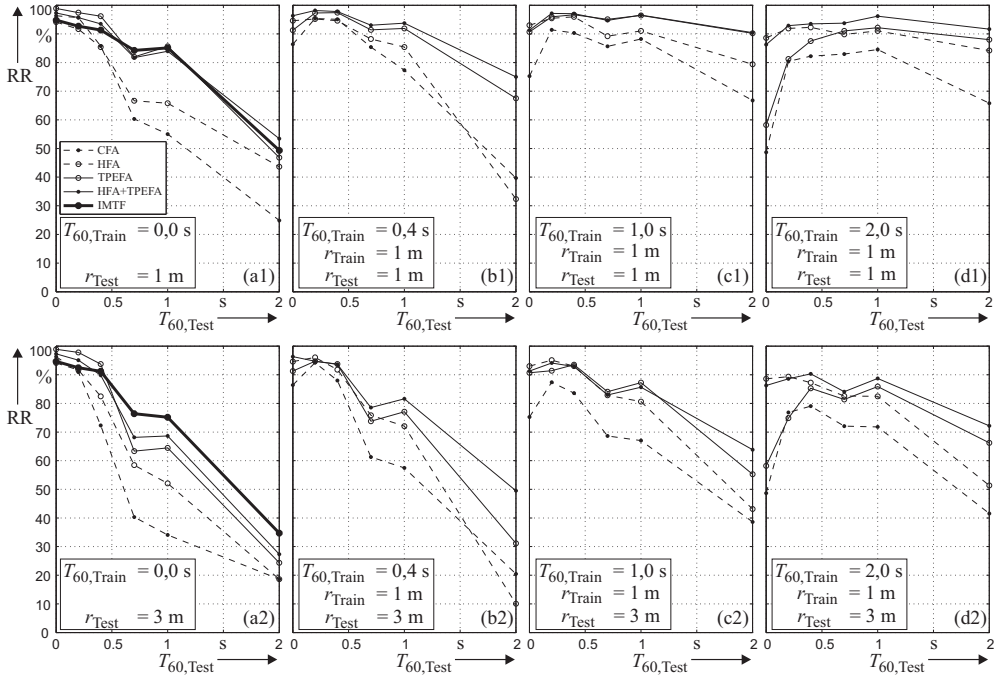

Abbildung 6.15 – Abhängigkeit der RR von T_{60}: Vergleich der Erkennungsergebnisse von CFA, HFA, TPEFA, HFA+TPEFA und IMTF. Oben, (a1) – (d1): SMD-Nahfeldbedingung $r_{\mathrm{Test}} = 1$ m. Unten, (a2) – (d2): SMD-Fernfeldbedingung $r_{\mathrm{Test}} = 3$ m. (a1), (a2): ungestörte Trainingsdaten (die IMTF-Methode wird nur in (a1) und (a2) dargestellt). (b1) – (d1), (b2) – (d2): verhallte Trainingsdaten ($T_{60,\mathrm{Train}} = (400; 1000; 2000)$ ms).

bungen erzielt. Hier wirkt die IMTF, da die benötigte Fernfeldbedingung greift. Die Steigerungen gegenüber TPEFA von ca. (5 ... 10) % korrespondieren in etwa mit den Ergebnissen in [LUA08]. Die geforderte Praxisbedingung RR = 90 % wird auch mit der IMTF-Methode mit Abstand nicht erreicht.

Verhallte Trainingsdaten (Abbildung 6.15 (b1) – (d1), (b2) – (d2))

TPEFA: Gegenüber dem ungestörten Training kann durch ein verhalltes Training die Leistung von TPEFA nochmals erheblich gesteigert werden, die Methode schneidet deutlich besser ab als CFA. Gute Ergebnisse werden, wie bereits für CFA und HFA, für die Trainingsbedingung $T_{60,\mathrm{Train}} = 1,0$ s erzielt. Im Nahfeld (c1) wird die Praxisbedingung RR > 90 % erfüllt, bemerkenswerterweise sogar für das Treppenhaus $T_{60,\mathrm{Test}} = 2,0$ s. Im Fernfeld (c2) kommt TPEFA dieser Bedingung im relevanten Be-

reich $T_{60,\text{Test}} = (0,3 \ \ldots \ 1,0)$ s mit ca. 85 % bereits sehr nahe. Es ist zu erkennen, dass TPEFA besonders gute Ergebnisse erzielt, wenn der Bereich der Testbedingungen in etwa mit der Trainingsbedingung übereinstimmt. Wie für CFA gilt, dass in den jeweils anderen Bereichen die Erkennungsrate schwächer wird (gut zu beobachten in (d1)). Hier besitzt TPEFA einen Nachteil gegenüber HFA, das die Erkennungsrate stabil verbessert.

HFA+TPEFA: Der Vorteil von HFA, die Erkennungsrate nicht nur für die benutzte Trainingsbedingung, sondern für einen breiten Bereich an Trainingsbedingungen zu verbessern (vgl. Abschnitt 6.4), steigert bei HFA+TPEFA nochmals die Leistung von TPEFA. Dieser Effekt ist besonders gut in (d1) und (d2) zu beobachten, man erkennt die Stabilität (und auch eine kleine Steigerung) gegenüber TPEFA.

IMTF: Der Sinn der IMTF-Methode ist es, einen verhallten TPE zu enthallen und daraus den sauberen TPE wieder herzustellen (Restauration). Ein verhalltes Training ist somit nicht sinnvoll.

Allgemein kann gesagt werden, dass verhalltes Training für alle Front-End-Methoden die Erkennungsraten erheblich steigert. Die besten Ergebnisse werden für HFA+TPEFA bei $T_{60,\text{Train}} = 1,0$ s und $r_{\text{Train}} = 1$ m erzielt ((c1) und (c2)).

6.5.2.2 Abhängigkeit der RR vom SMD

Die Ergebnisse dieser Experimente werden in Abbildung 6.16 zusammengefasst. Analog zu Abschnitt 6.4.2 wird mit sauberen und mit verhallten Sprachdaten gearbeitet. Es wird wieder die SMART-Room-Umgebung ($T_{60,\text{Test}} = 0{,}7$ s) genutzt. Um Aufwand zu sparen, wird hier im Unterschied zu Abbildung 6.7 nicht mehr mit allen zur Verfügung stehenden SMDs, sondern nur mit den ausgewählten SMDs $r_{\text{Train}} = (60; 140; 280)$ cm trainiert. Es entsteht deshalb keine 3D-Graphik wie in Abbildung 6.7, sondern die vier 2D-Diagramme. Zum Vergleich werden wieder die Ergebnisse für CFA und HFA mit abgebildet.

Ungestörte Trainingsdaten (Abbildung 6.16 (a))

TPEFA: Obwohl die TPEFA-Methode deutlich bessere Ergebnisse erzielt als CFA, erkennt man auch für TPEFA, dass bereits nach einigen cm SMD die Erkennungsrate stark einbricht. Bis zum SMD von etwa 1 m sind die Ergebnisse für TPEFA noch schlechter als für HFA. Das wird damit erklärt, dass HFA bei einem sehr großen Anteil von Direktschall besonders gut arbeitet und die betreffenden störenden Anteile eliminiert. Die Folge ist, dass die Erkennungsrate bei HFA in diesem SMD-Bereich relativ konstant bleibt, allerdings wegen der bereits genannten Eliminierung von Informationen schlechter ist als bei sauberen CFA- oder TPEFA-Ergebnissen. Steigt der SMD in Richtung Hallradius, schneiden HFA und TPEFA in etwa gleich ab. Dieses Verhalten korrespondiert nicht vollständig mit dem Verhalten in Abbildung 6.15 (a1) bzw. (a2).

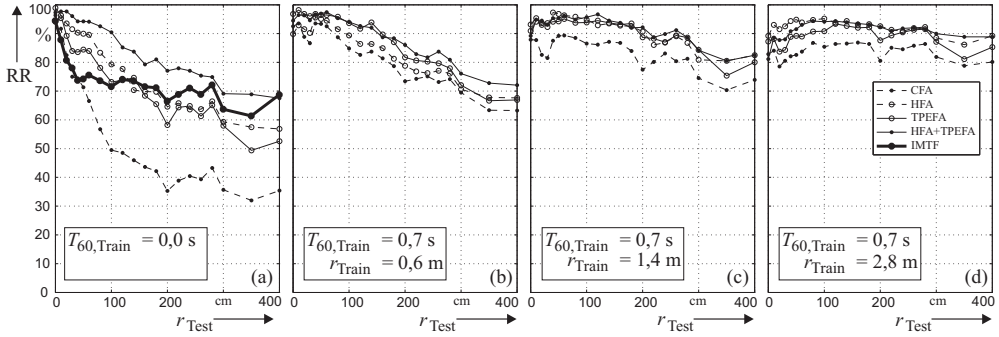

Abbildung 6.16 – Abhängigkeit der RR vom SMD im SMART-Room: Vergleich der Erkennungsergebnisse von CFA, HFA, TPEFA, HFA+TPEFA und IMTF. (a): ungestörte Trainingsdaten (die IMTF-Methode wird nur in (a) dargestellt). (b) – (d): verhallte Trainingsdaten ($r_{\mathrm{Train}} = (60; 140; 280)$ cm).

Es wird deshalb geschlussfolgert, dass die Algorithmen auf Individualitäten der einzelnen Räume in geringen Grenzen unterschiedlich reagieren. Dass reale Räume trotz gleichem T_{60} Unterschiede aufweisen, wie z. B. besondere Eigenmoden oder besonders starke diskrete Reflexionen, wurde ja in dieser Arbeit bereits mehrfach betont.

HFA+TPEFA: Für diesen Raum erkennt man besonders gut, dass die unterschiedlichen Vorteile von HFA und TPEFA sich gegenseitig ergänzen, wodurch die Erkennungsrate jeweils noch einmal um ca. 10 % gesteigert werden kann. Die Gesamtsteigerung im Vergleich zu CFA beträgt somit beachtliche 40 %.

IMTF: In dieser Graphik erkennt man am Vergleich TPEFA vs. IMTF deutlich, wie für das Nahfeld zunächst das Merkmalmuster durch die für das Fernfeld konzipierte IMTF gestört wird (Ergebnisse für $r_{\mathrm{Test}} < 1$ m). Wird der SMD in Richtung Hallradius und Fernfeld erhöht, erkennt man, wie die IMTF zunächst die Ergebnisse von TPEFA erreicht ($r_{\mathrm{Test}} = (1 \ldots 1,4)$ m) und später verbessert ($r_{\mathrm{Test}} > 1,4$ m). Man hat somit einen Nachweis, dass die IMTF-Methode tatsächlich nur im Fernfeld gewinnbringend arbeitet. Jedoch gilt auch hier, dass die geforderte Praxisbedingung (RR > 90 %) bei Weitem nicht erreicht wird.

Verhallte Trainingsdaten (Abbildung 6.16 (b) – (d))

TPEFA: Wie erwartet steigert das verhallte Training die Erkennungsrate erheblich. Im Falle von TPEFA (wie auch CFA) erkennt man wieder, dass die Testbedingungen (hier SMDs), die im Bereich der Trainingsbedingung liegen, gut erkannt werden. Liegen sie außerhalb, verschlechtert sich die Erkennungsrate (Beispiel: (d) $r_{\mathrm{Test}} < 1$ m). Im Vergleich dazu hält HFA die Erkennungsrate in (d) auch für kleine SMDs konstant.

HFA+TPEFA: Die Kombination von HFA+TPEFA erzielt fast ausschließlich die

besten Ergebnisse. TPEFA wird wieder durch die zusätzlichen Vorteile von HFA gesteigert. Besonders gute Ergebnisse werden beim SMD $r_{\text{Train}} = 2,8$ m erreicht (d). Allerdings schneidet HFA+TPEFA durch das ungünstige Nahfeldverhalten von TPEFA bei $r_{\text{Train}} = 2,8$ m im Nahfeld $r_{\text{Test}} < 1$ m schlechter ab als HFA, das hier bereits eine zufriedenstellende Erkennungsrate (RR > 90 %) für alle Testbedingungen erreicht (vgl. auch Abbildung 6.7).

IMTF: Wie oben festgestellt, ist verhalltes Training hier nicht sinnvoll.

6.5.2.3 Vergleich mit einem DSB für ungestörte Trainingsdaten

Da alle vorausgegangenen Experimente einkanalige Eingangssignale hatten, war es zunächst nötig, neue RIRs zu messen, um eine mehrkanalige DSB-Funktionalität zu erzeugen. Die bisherigen Ergebnisse sind somit nicht eindeutig mit den DSB-Ergebnissen in Abbildung 6.17 vergleichbar. Zur Messung wurden vier Räume ausgewählt (Tonstudio $T_{60} = 250$ ms, Wohnzimmer $T_{60} = 450$ ms, Büroraum $T_{60} = 700$ ms, Computerlabor $T_{60} = 950$ ms). In jedem Raum wurden die RIRs für mehrere SMDs gemessen ($r = (0,2;\ 0,4;\ \ldots;\ 4)$ m). Bei der Messung einer Sprecher-Mikrofon-Anordnung wurden die vier RIRs $h_m(k)$ simultan aufgenommen. Für die einkanaligen Methoden CFA und HFA wird jeweils $h_2(k)$ benutzt (vgl. Abbildung 6.14 (b), linkes mittleres Mikrofon). Um die DSB-Funktionalität zu simulieren, werden die vier RIRs $h_1(k) \ldots h_4(k)$ wie in Gleichung (6.31) addiert, um letztlich eine einzelne RIR $h_{\text{DSB}}(k)$ zu erhalten. Anschließend erfolgen die Faltungen von $h_2(k)$ bzw. $h_{\text{DSB}}(k)$ mit den Sprachsignalen des APOLLO-Korpus, analog zu den vorangegangenen Experimenten.

Die Experimente wurden nur für unverhalltes Training durchgeführt. Abbildung 6.17 zeigt die Ergebnisse.

CFA: Ähnlich wie in den vorigen Experimenten fällt RR für CFA mit steigender T_{60} (Teilabbildungen von links nach rechts). Zusätzlich bietet diese Graphik den direkten Zusammenhang zwischen T_{60} und SMD. Man erkennt wie zuvor in Abbildung 6.16 (a) (SMART-Room), dass RR für einen steigenden SMD fällt. Die Umgebung in Abbildung 6.17 (c) kann in etwa mit den Ergebnissen im SMART-Room verglichen werden. Am Verhalten der Erkennungsraten wird deutlich der Übergang zwischen Nah- und Fernfeld erkennbar.

DSB: Für alle halligen Räume ((b) bis (d)) erkennt man eine leichte Steigerung der Erkennungsrate von etwa (5 ... 10) % durch den Einfluss des DSBs. Damit ist die zuvor getroffene Annahme der begrenzten Wirkung des DSBs für die Spracherkennung nachgewiesen. Im extremen Nahfeld $r_{\text{Test}} = 20$ cm erkennt man in einigen Fällen eine leichte Verschlechterung der Erkennungsrate gegenüber dem größeren SMD $r_{\text{Test}} = 0,4$ m, die damit erklärt werden kann, dass die Schallwellen hier nicht mehr als ebene Wellen betrachtet werden können (Kugelwelle mit geringem Radius). Die Verzögerungen κ_τ müssten für diese geringe Entfernung angepasst werden.

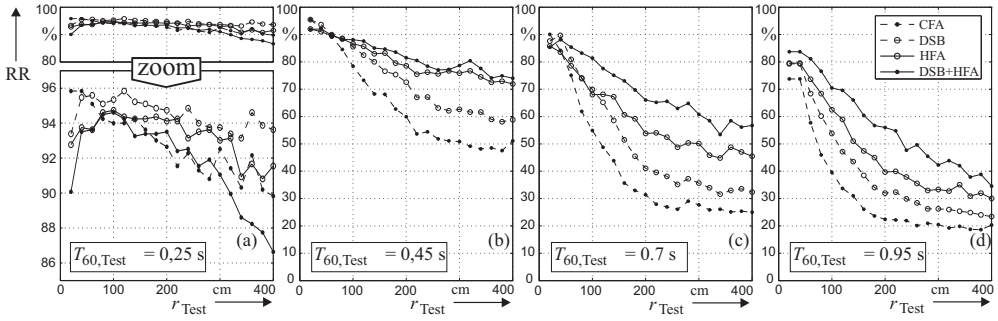

Abbildung 6.17 – Gemessene Erkennungsraten für die vier Front-Ends CFA, DSB, HFA und DSB+HFA abhängig vom SMD r_{Test}. Die Teilabbildungen (a) bis (d) zeigen die Ergebnisse für die vier Räume in der Reihenfolge aufsteigender Nachhallzeit $T_{60} = (250; 450; 700; 950)$ ms. Teilabbildung (a) ist im relevanten Bereich vergrößert worden, um Einzelheiten besser darzustellen.

HFA: HFA kann im Vergleich zum DSB für hallige Bedingungen deutliche Steigerungen erzielen. Sind die Bedingungen weniger hallig, so verringert sich die Erkennungsrate im Vergleich zu CFA wie schon in vorhergehenden Experimenten geringfügig (z. B. in Teilabbildung (a) bis (c) bei $r_{\mathrm{Test}} < (0,5 \ \ldots \ 1)$ m), weil hier durch HFA wieder teilweise nützliche Informationen eliminiert werden. Das bessere Abschneiden von HFA im Vergleich zum DSB wird damit erklärt, dass HFA speziell für die spezifischen Eigenschaften einer Hallstörung ausgelegt ist, wogegen der DSB gerade die weniger störenden (hochfrequenten) Hallanteile unterdrückt, die besonders störenden (tieffrequenten) aber im Signal belässt.

DSB+HFA: Kombiniert man DSB und HFA, können sich die verbessernden Effekte beider Methoden addieren, sodass teilweise Gesamtsteigerungen von bis zu 35 % Erkennungsrate im Fernfeld erreicht werden können.

Im Rahmen der vorliegenden Arbeit wurden keine Experimente für verhalltes Training für die Experimente im Zusammenhang mit DSBs durchgeführt. Es kann jedoch eine ähnliche Steigerung der Erkennungsrate wie bei den Experimenten in Abbildung 6.15 und 6.16 angenommen werden.

6.6 Zusammenfassung der Erkenntnisse

Dieses Kapitel stellt die im Rahmen der vorliegenden Arbeit entwickelte Methode HFA vor. Sie basiert auf den drei Ideen, dass harmonische Komponenten im Spektrum als ungestört, tiefe Frequenzen in stimmlosen Abschnitten besonders störend sowie hochfrequente Hallkomponenten als harmlos angenommen werden können. Diese Ide-

en führen zu einer Methode, die störende Merkmale unterdrückt und die Spracherkennung nur auf nicht störende Merkmale ausrichtet. HFA besteht aus vier Submodulen, für die in einer Optimierung sechs feste Parametereinstellungen gefunden werden. Alle Submodule arbeiten unter Hallbedingungen teilweise fehlerhaft. Es kann jedoch dargestellt werden, dass die Fehler im HFA-Ansatz einkalkuliert werden, sodass sie sich nicht negativ auswirken.

Die HFA-Methode wird von vornherein unter Berücksichtigung von praktischen Einsatzbedingungen konzipiert. Sie benötigt dadurch nur geringe zusätzliche Rechenleistung. Sie besitzt keine Adaption und kann damit in Echtzeit arbeiten.

Die HFA-Methode wird umfangreich evaluiert. Zunächst erfolgt eine Evaluation im Vergleich mit CFA, um die verbessernde Wirkung von HFA für die Spracherkennung zu testen. Für die Experimente werden verschiedene Hallbedingungen systematisch variiert. Dabei wird in die Abhängigkeit der Erkennungsrate von T_{60} sowie vom SMD unterschieden. Es kann gezeigt werden, dass HFA die Erkennungsrate für hallige Umgebungen steigern kann. Für weniger hallige Signale verschlechtert sich die Erkennungsrate geringfügig im akzeptablen Bereich, da HFA einige für die Erkennung relevante Informationen eliminiert. Die alleinige Verbesserung durch HFA erreicht allerdings noch nicht die Praxisbedingung RR > 90 %. Diese wird durch zusätzliches Verhallen der Trainingsdaten erreicht. Hier zeigt HFA bessere Erkennungsraten als CFA, die der Praxisbedingung RR > 90 % genügen. Dabei kann im Vergleich zu CFA insbesondere die Stabilität gegen veränderliche Hallbedingungen betont werden.

Um die Leistungsfähigkeit von HFA in den aktuellen Stand der Technik einzuordnen, werden Experimente mit drei weiteren aus der Literatur bekannten Methoden durchgeführt. Dies sind TPEFA, IMTF-basierte Enthallung sowie DSB. Zusätzlich werden die Kombinationen HFA+TPEFA und DSB+HFA getestet. Für unverhalltes Training zeigt sich, dass TPEFA bedeutende Steigerungen der Erkennungsrate erzielt, die die Leistungsfähigkeit von HFA teilweise übertrifft. Die IMTF-Methode, die auf der TPEFA-Methode basiert und die generierten TPEs enthallen soll, zeigt prinzipbedingt nur im halligen Fernfeld eine steigernde Wirkung. Im Nahfeld sowie bei gering halligen Daten fügt die Methode teilweise sogar Störungen ein, da sie für diese Bedingungen nicht ausgelegt ist. Die Erkennungsrate verschlechtert sich geringfügig. Der DSB kann die Erkennungsrate ebenfalls erwartungsgemäß nur geringfügig steigern. Die Zusammenschaltungen HFA+TPEFA sowie DSB+HFA können die Vorteile der beiden betreffenden Methoden kombinieren und erzielen dadurch eine weitere Steigerung. Trotz erheblicher Steigerungen (teilweise bis zu 40 %) wird für unverhalltes Training die Praxisbedingung RR > 90 % nur in wenig hallenden Umgebungen erreicht. Deshalb werden zusätzlich für die Methoden CFA, HFA, TPEFA und HFA+TPEFA Experimente mit verhalltem Training durchgeführt. Im Ergebnis sind deutliche Steigerungen für die betreffenden Methoden festzustellen. Dabei wird bei den relevanten Hallbedingungen des Wohn- und Büroumfeldes die Praxisbedingung RR > 90 % beim Training mit der Bedingung $T_{60,\mathrm{Train}} = 1$ s, $r_{\mathrm{Train}} = 100$ cm meist erreicht. Bei diesen Experi-

menten wird festgestellt, dass Erkennungsergebnisse für Testdaten, die der trainierten Umgebungsbedingung entsprechen, durch CFA und TPEFA bevorzugt, andere Testbedingungen jedoch schlechter erkannt werden. Bei HFA tritt dieser Effekt nicht (bzw. wesentlich geringer) auf, wodurch die Methode in Kombination mit verhalltem Training nicht sensibel auf Veränderungen der akustischen Umgebung reagiert (vgl. Praxisanforderungen). Gleiches gilt für die Front-End-Methode HFA+TPEFA, die durch die Kombination der Vorteile von HFA (stabile Steigerung der Erkennungsrate über veränderliche Hallbedingungen) und TPEFA (hohe Erkennungsraten zur passenden Trainingsbedingung) die besten Ergebnisse aller Methoden erzielt.

7 F_0-Detektion und VUD unter verhallten Umgebungsbedingungen

7.1 Motivation und Überblick

Die automatische Messung der Grundfrequenz F_0 menschlicher Sprache (kurz: F_0-Detektion) sowie die Stimmhaft-Stimmlos-Entscheidung (engl.: VUD – Voiced Unvoiced Decision) sind zwei traditionelle und miteinander gekoppelte Teilgebiete der Sprachsignalverarbeitung. Obwohl diese Arbeit sich mit Spracherkennung befasst, ist die kurze Behandlung beider Gebiete erforderlich. Beide Technologien müssen für die zu bewältigende Aufgabe der verhallten Spracherkennung innerhalb der in Abschnitt 6 eingeführten Methode HFA unter Hallbedingungen arbeiten. Deshalb werden in diesem Kapitel die wichtigsten bestehenden Verfahren zur F_0-Detektion und VUD unter Hallbedingungen evaluiert.

Bereits frühe Untersuchungen vergleichen mehrere F_0-Detektionsverfahren unter sauberen Bedingungen (z. B. [RCRM76]). Den umfassendsten Überblick enthält [Hes83] bzw. etwas neuer [Hes92, dCK01]. Studien der letzten 10 bis 20 Jahre untersuchen auch das Verhalten der Methoden unter Geräuschbedingungen und arbeiten an der Verbesserung der Geräuschrobustheit (guter Überblick in [Hes08] oder [UH08c]). Wie bereits festgestellt wurde, kann die Hallstörung mit der Geräuschstörung nicht verglichen werden, da beide völlig unterschiedlich wirken. Demnach ist das Heranziehen derartiger Studien hier nicht bzw. nur bedingt relevant.

Das Verhalten von F_0-Detektionsverfahren unter Hallbedingungen ist zum aktuellen Zeitpunkt nicht untersucht. Als bislang einzige Studien existieren die Arbeiten von Unoki et al. [UH08c, UHI08]. [UH08c] evaluiert das Verhalten von zwölf F_0-Detektionsmethoden unter Verhallung durch künstlich generierte RIRs nach Gleichung (3.81). [UHI08] erweitert die Studie um Verhallungen mit realen RIRs. Diese wurden jedoch hauptsächlich in großen Räumen oder Hallen, wie z. B. Konzertsälen, gemessen und besitzen vorwiegend SMDs von mehreren Metern. Deshalb beschreiben beide Studien nur das Verhalten im Fernfeld. Hier wird jedoch speziell auch auf das Nahfeld und auf das Verhalten in Abhängigkeit vom SMD eingegangen. Ein weiterer Unterschied zu [UH08c] ist die geschlechterspezifische Untersuchung der F_0-Detektionsmethoden, was ebenfalls zu neuen Erkenntnissen führt.

Die vorliegende Untersuchung schließt an die Arbeiten von [UH08c] an und ergänzt sie

durch Untersuchung von Abhängigkeiten bei realen Hallbedingungen aus dem Wohn- und Büroumfeld. Dabei folgt sie dem Schema der Evaluationen, wie es in Abschnitt 6.4 bereits für die Spracherkennung vorgestellt wurde. Die Untersuchungen haben zwei Ziele:

1. Feststellung des Verhaltens von F_0-Detektions- und VUD-Verfahren unter realen Hallbedingungen im Wohn- und Büroumfeld.
2. Auswahl je eines passenden Verfahrens für die HFA-Methode unter Berücksichtigung der Kriterien für anwendungsfähige, integrierte Spracherkennung (vgl. Abschnitt 1.2).

Zusätzlich liefern die Untersuchungen Aussagen für weitere Aufgaben der Sprachsignalverarbeitung, welche auf F_0-Detektion bzw. VUD angewiesen sind (z. B. Sprachsynthese, spezielle Geräuschunterdrückungs- und Enthallungsverfahren, Sprach-Pause-Detektion[1]).

7.2 Technologische Einführung

7.2.1 Klassen von Verfahren

Die Thematik der Bestimmung menschlicher Stimme/Sprache lässt sich in vier Abstraktionsebenen einteilen (Schwierigkeitsgrad steigend):

1. **VAD – Voice Activity Detection**[2] beschreibt eine Klasse von Methoden, deren Aufgabe es ist, das Vorhandensein menschlicher Sprachsignale festzustellen. Das Ergebnis ist eine binäre Sprach-Pause-Funktion $VAD(t)$.
2. **VUD – Voiced Unvoiced Decision**[3] beschreibt eine Klasse von Methoden, die das Vorhandensein stimmhafter bzw. stimmloser Sprachesignale detektieren. Das Ergebnis ist eine binäre Stimmhaft-Stimmlos-Entscheidung $VUD(t)$ (v/u - voiced/unvoiced), kann allerdings auch durch vier Zustände (1. stimmhaft, 2. stimmlos, 3. stimmhaft und stimmlos (Mischanregung), 4. Pause) repräsentiert werden [Hes08].
3. **PDA – Pitch Determination Algorithm**[4] beschreibt ein Verfahren zur F_0-Detektion. Das Ergebnis ist eine geschätzte F_0-Zeit-Funktion $F_{0,\mathrm{Est}}(t)$ (Index

[1] Engl. Fachbegriff: Voice Activity Detection (VAD)

[2] Der Begriff ist fest etabliert, obwohl er etwas unglücklich gewählt ist, da er eigentlich eher die Aufgabe eines VUD beschreibt. Speech Activity Detection (SAD) oder Speech Pause Decision (SPD) wären günstigere Begriffe gewesen.

[3] Der Begriff erscheint so z. B. in [PLLH08, RLM97]. Es werden außerdem die Begriffe Voiced Unvoiced Detection [KHK06, MHMH07], Voicing Determination und zugehörig Voicing Determination Algorithm (VDA) u. a. nach [Hes92, Hes08] benutzt.

[4] Der Begriff wird in [Hes83, Hes08] gebraucht. Weitere übliche Begriffe sind Grundfrequenzdetektion [VHH98], oder engl. Fundamental Frequency Estimation [UH08c], F_0-Estimation, Pitch Extraction [Ger03].

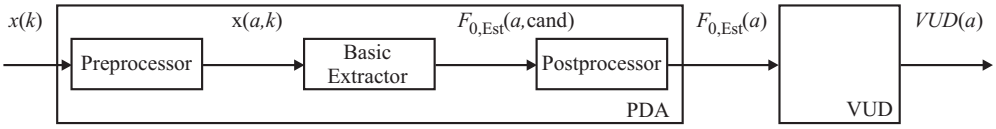

Abbildung 7.1 – Klassischer struktureller Aufbau von F_0-Detektionsmethoden gefolgt von einem VUD. Graphik und Begrifflichkeiten angelehnt an [Hes83].

Est für Estimation).

4. **PMA – Pitch Marking Algorithm** beschreibt eine Klasse von Verfahren, die im Zeitbereich den festgelegten[5] Anfang einer Grundperiode markieren. Das Ergebnis ist eine Folge von Periodenmarken $PM(i)$.

In der vorliegenden Arbeit werden von diesen vier nur die Klassen PDA und VUD behandelt. Für beide sind in der Vergangenheit verschiedene Ansätze entwickelt worden. Einen Überblick über die Verfahren beschreibt [Hes08]. Der allgemeine Aufbau wird in den folgenden beiden Abschnitten dargestellt. Einzelne Verfahren werden in den Abschnitten 7.4 und 7.6 kurz vorgestellt.

7.2.2 Allgemeiner Aufbau von PDA-Verfahren

Der klassische strukturelle Aufbau von Methoden der F_0-Detektion (Abbildung 7.1) kann für die meisten Verfahren angewandt werden. Es sind drei Blöcke zu erkennen:

- **Preprocessor** Das eintreffende Sprachsignal $x(k)$ wird ähnlich wie bei einem Spracherkenner in eine Folge von Kurzzeitsignalabschnitten $\boldsymbol{x}(a)$ (Frames, Frameindex a) überführt. In der Literatur der F_0-Detektionsverfahren wird dieses Element Preprocessor genannt. Der Aufbau des Preprocessors ist identisch mit dem Framing bei der Spracherkennung (vgl. Abschnitt 2.2.1 Framing). Die Analyseframelänge sollte so kurz wie möglich gewählt werden, da die Grundfrequenz nicht konstant ist. Dennoch sollte sie so groß gestaltet werden, dass mindestens drei Grundperioden enthalten sind [Hes08]. Beides steht im Widerspruch zueinander und kann im Falle von sehr schnellen Änderungen der Grundfrequenz zu prinzipbedingten Fehlern führen. Laut (4.5) wäre dann die kleinste Analysefensterlänge von $3 \cdot T_{0,\max} = 42,86$ ms nötig (in den Experimenten wurde 40 ms benutzt).
- **Basic-Extraktor** Der sich anschließende Basic-Extraktor ist die eigentliche F_0-Detektionsmethode. In der Vergangenheit wurden eine ganze Reihe von Algorithmen entwickelt, die sich in ihrer Arbeitsweise stark unterscheiden. Der Basic-

[5] Mathematisch gesehen haben Perioden keinen Anfang. Deshalb wird ein typischer wiederkehrender Zeitpunkt als Anfang bzw. Periodenmarke definiert. Oft wird dafür der Punkt des globalen Minimums einer Periode gewählt [HKKP06].

Extraktor generiert eine frameweise Folge von F_0-Schätzungen $F_{0,\mathrm{Est}}(a, \mathrm{cand})$, die, je nach Ausführung, einen bis mehrere F_0-Kandidaten (Kandidatenindex cand) hervorbringt. Hier befindet sich noch keine VUD-Entscheidung, d. h., zu jedem Frame, also auch zu stimmlosen, werden F_0-Kandidaten vom Algorithmus vorgeschlagen.

- **Postprocessor** Der Postprocessor hat die Aufgabe, die richtige Kandidatenfolge zu spezifizieren und daraus das Endergebnis $F_{0,\mathrm{Est}}(a)$ der F_0-Detektion zu ermitteln. Glättungs- oder Impulsunterdrückungsverfahren können ebenfalls implementiert werden. Aus der Anzahl der Kandidaten wird einer ausgewählt. Im Postprocessor kann auch die VUD-Entscheidung getroffen werden. Im Falle von stimmlosen Sprachframes a_u würde man dann einen geeigneten Wert für $F_0(a_u)$ setzen, vorzugsweise $F_0(a_u) = 0$ Hz.

Es können verschieden Methoden des Basic-Extraktors mit verschiedenen Methoden des Postprocessors kombiniert werden. Deshalb werden in dieser Arbeit beide unabhängig voneinander getestet.

7.2.3 Allgemeiner Aufbau von VUD-Verfahren

Die Grobstruktur von VUD-Verfahren ähnelt der von F_0-Detektionsverfahren, wobei sich noch ein VUD-Block an das Blockschaltbild anschließt (Abbildung 7.1). Allerdings ist diese Struktur nur eine abstrakte Möglichkeit der Verallgemeinerung. Es ist möglich, dass ein VUD auf einer vorhergehenden F_0-Detektion basiert oder auch nicht. Im letzteren Fall würden die Blöcke Basic-Extraktor und Postprocessor entfallen. Oft ist jedoch der erste Fall – ein VUD basiert auf einer vorhergehenden F_0-Detektion – nötig. Im Umkehrschluss darf auch gesagt werden, dass die VUD-Entscheidung zur F_0-Detektion gehört, da sie in stimmlosen Abschnitten fehlgeschätzte F_0 korrigiert. Da dies auch eine Aufgabe des Postprocessings ist, darf weiterhin gesagt werden, dass Postprocessing-Verfahren die Aufgabe eines VUDs übernehmen können. In diesem Fall würden der Postprocessing-Block und der VUD-Block in einem Blockschaltbild verschmelzen. Zunächst soll das vollständige Blockschaltbild aus Abbildung 7.1 gelten.

7.3 Evaluationsumgebung

Ein entsprechendes Verfahren der F_0-Detektion bzw. VUD wird auf sein Störverhalten (hier Hall) hin evaluiert, indem es auf eine signifikante Anzahl von Sprachdaten bei einer bestimmten (Hall-)Bedingung angewendet wird. Dabei entstehen die vom Verfahren ermittelten (geschätzten, engl.: estimated, Index Est) Werte für $F_{0,\mathrm{Est}}(a)$ bzw. $VUD_{\mathrm{Est}}(a)$. Aus dem Vergleich mit den Referenzwerten $F_{0,\mathrm{Ref}}(a)$ bzw. $VUD_{\mathrm{Ref}}(a)$ ermittelt eine Auswerteeinheit einen oder mehrere Evaluationsparameter. Diese Prozedur wird für definiert veränderte Störbedingungen (Hall) durchgeführt, sodass die

Evaluationsparameter eine Abhängigkeit von einem Störparameter (z. B. T_{60}, SMD) ergeben. Die Abhängigkeit soll möglichst graphisch aufbereitet werden. Im Folgenden werden die benötigten Elemente eines solchen Testsystems beschrieben:

- Sprachdatenbasen (Abschnitt 7.3.1),
- Referenzdaten (Abschnitt 7.3.2),
- Störung durch definierte Verhallung (Abschnitt 7.3.3) und
- Evaluationsparameter (Abschnitt 7.3.4).

7.3.1 Sprachdatenbasen

Um entsprechende Verfahren der F_0-Detektion bzw. der VUD evaluieren zu können, benötigt man ähnlich wie bei der Spracherkennung eine Sprachdatenbasis, auf deren Sprachsignale die Verfahren angewandt werden. Die Sprachdatenbasis muss zusätzlich zu den Sprachdaten zugehörige Referenz-F_0-Daten enthalten, mit denen die Ergebnisse der F_0-Detektions- bzw. VUD-Verfahren verglichen werden.

Für die hier vorgestellten Experimente wurden drei Datenbasen benutzt:

- **(i) Japanische Datenbasis** - In den Experimenten von Unoki et al. [UH08c, UHI08], an die hier angeschlossen werden soll, wurde eine japanische Sprachdatenbasis verwendet, die von Atke et al. [AIK+00] erstellt wurde. Sie besteht aus 30 kurzen japanischen Sätzen, gesprochen von jeweils 14 männlichen und 14 weiblichen Sprechern, die zwischen 22 und 36 Jahren alt waren. Insgesamt entstehen dabei 840 Äußerungen bestehend aus 36 min Sprache (davon 21 min stimmhaft) mit $f_s = 16$ kHz. Die Referenzdaten wurden aus einer simultan aufgenommenen Elektroglottograph-Aufzeichnung (EGG) gebildet. Dabei wurde das TEMPO-Verfahren (vgl. Abschnitt 7.4.1) auf das EGG-Signal angewendet.
- **(ii) Englische Datenbasis (klein)** - Die hier vorgestellten Experimente zur F_0-Detektion wurden zunächst mit einer kleinen englischen (Großbritannien) Sprachdatenbasis durchgeführt. Sie besteht aus einem Subset der im Rahmen des EU-Projektes TC-STAR (Technology and Corpora for Speech Translation) [BHT+04] aufgenommen Datenbasis und enthält Äußerungen von einem männlichen und einem weiblichen Sprecher, welche die gleichen 19 Sätze sprechen (insgesamt 38 Äußerungen bestehend aus 275/263 s Sprache, $f_s = 16$ kHz). Die Referenzdaten wurden im Rahmen von [HJ07] durch einen zweistufigen Prozess erzeugt. Der erste Schritt bestand aus einem hybriden Prozess nach Hussein et al. [HJ07], der die Ergebnisse eines EGG-Signals und eines Periodenmarkierungsalgorithmus (Complex Harmonic Filter Analysis (CHFA) [Duc06]) miteinander kombiniert. Im zweiten Schritt wurden die ermittelten Periodenmarken von Hand korrigiert. Als Referenz wurde letztlich ein Satz von Periodenmarken zur Verfügung gestellt.

- **(iii) ECESS-Datenbasis** - Im späteren Verlauf der Arbeit wurde mit einer größeren Datenbasis gearbeitet, die in den hier vorgestellten Experimenten im Wesentlichen für die Untersuchung der VUD-Methoden benutzt wurde. Sie wird vom European Center of Excellence in Speech Synthesis (ECESS) [ECESS] zur Verfügung gestellt und enthält von elf weiblichen und elf männlichen englischen sowie elf weiblichen und elf männlichen deutschen Muttersprachlern gesprochenen Sätze. Die Referenzdaten wurden von ECESS in Form von Periodenmarken zur Verfügung gestellt.

7.3.2 Erstellung der Referenzdaten

Im vorigen Abschnitt wurde bereits die Aufbereitungsform der zu den entsprechenden Datenbasen gelieferten Referenzdaten genannt. Die Referenzdaten für F_0 bzw. VUD unterscheiden sich anhand ihrer Aufgabenstellung im Wertebereich. Während sich der Wertebereich für $F_{0,\mathrm{Ref}}(a)$ stufenlos[6] in den Grenzen von $F_{0,\min}$ und $F_{0,\max}$ bewegt (vgl. (4.5)), ist $VUD_{\mathrm{Ref}}(a)$ ein Binärwert (v/u – voiced/unvoiced).

Da hier beide Klassen von Verfahren auf gleichen Datenbasen getestet werden, ist $VUD_{\mathrm{Ref}}(a)$ bereits in $F_{0,\mathrm{Ref}}(a)$ enthalten.

$$\begin{array}{lllll} VUD_{\mathrm{Ref}}(a) = v & ; & F_{0,\min} & \leq F_{0,\mathrm{Ref}}(a) & \leq F_{0,\max}, \\ VUD_{\mathrm{Ref}}(a) = u & ; & F_{0,\mathrm{Ref}}(a) & = 0. & \end{array} \tag{7.1}$$

Die Referenzkurve der Sprachdatenbasis **(i)** lag bereits als F_0-Zeit-Funktion mit $f_s = 16$ kHz vor. Die Referenz der beiden anderen Datenbasen **(ii)** und **(iii)** war in Form von Periodenmarken gegeben. Daraus wurde die zugehörige F_0-Zeit-Funktion berechnet.

7.3.3 Hallbedingungen

Die ausgewählten PDAs und VUDs wurden unter den gleichen Hallbedingungen getestet wie in den Spracherkennungsexperimenten in Kapitel 6. Dazu wurden die Korpora **(i)** bis **(iii)** mit den entsprechenden RIRs gefaltet. Wichtig dabei ist, dass die RIRs exakt mit dem Direktschallimpuls beginnen (die Delay-Phase muss manuell abgeschnitten werden), da sonst die Referenz-PMs (und damit die Referenz-F_0) zeitlich verschoben sind. Es entstehen die vier Abhängigkeiten der Evaluationsparameter von

- (a) T_{60} bei Verhallungen mit künstlichen RIRs (beschreibt das Fernfeld),
- (b) SMD bei Verhallungen mit realen RIRs im SMART-Room ($T_{60} \approx 0.7$ s),
- (c) T_{60} bei Verhallungen mit realen RIRs mit Nahfeld-SMD ($r = 1$ m) und
- (d) T_{60} bei Verhallungen mit realen RIRs mit Fernfeld-SMD ($r = 3$ m).

[6] Für die Repräsentation in ganzzahligen Digitalwerten ist die Auflösung dabei hinreichend klein zu wählen. Im Allgemeinen reicht aber eine Repräsentation in 16-Bit Ganzzahlen aus.

Tabelle 7.1 – Typische Fehlerbereiche und daraus ermittelte Gütemaße bei PDA-Evaluationen. Der Name des Gütemaßes leitet sich aus den gezählten Klassifikationsevents ab: Fine Correct Rate (FCR), Fine Error Rate (FER), Gross Correct Rate (GCR), Gross Error Rate (GER).

Frame-Klassifikation	Bedingung	Eventzähler	Berechnung des Gütemaßes
Fine Correct	$\Delta_{F_0(a)} < \Delta_{F_0,\mathrm{FE}}(a)$	N_{FC}	$\mathrm{FCR} = \frac{N_{FC}}{N_{v,\mathrm{Ref}}}$
Fine Error	$\Delta_{F_0(a)} > \Delta_{F_0,\mathrm{FE}}(a)$	N_{FE}	$\mathrm{FER} = \frac{N_{FE}}{N_{v,\mathrm{Ref}}}$
Gross Correct	$\Delta_{F_0(a)} < \Delta_{F_0,\mathrm{GE}}(a)$	N_{GC}	$\mathrm{GCR} = \frac{N_{GC}}{N_{v,\mathrm{Ref}}}$
Gross Error	$\Delta_{F_0(a)} > \Delta_{F_0,\mathrm{GE}}(a)$	N_{GE}	$\mathrm{GER} = \frac{N_{GE}}{N_{v,\mathrm{Ref}}}$

Um einen Anschlusspunkt an die Experimente von [UH08c] zu schaffen, wurden die Experimente für PDAs, im Gegensatz zu den Spracherkennungsexperimenten, nicht nur für die realen, sondern auch für den künstlichen RIRs durchgeführt.

7.3.4 Evaluationsparameter

Die vom PDA bzw. VUD ermittelten Ergebnisse ($F_{0,\mathrm{Est}}(a)$, $VUD_{\mathrm{Est}}(a)$) werden in der Evaluation mit den Referenz-F_0-Daten $F_{0,\mathrm{Ref}}(a)$ verglichen. In der Literatur sind bereits traditionelle Evaluationsparameter etabliert.

7.3.4.1 PDA-Evaluationsparameter

Maße für die Güte von PDAs werden aus dem frameweisen absoluten Fehler der Messung berechnet

$$\Delta_{F_0}(a_{v,\mathrm{Ref}}) = \left|F_{0,\mathrm{Est}}(a_{v,\mathrm{Ref}}) - F_{0,\mathrm{Ref}}(a_{v,\mathrm{Ref}})\right|. \tag{7.2}$$

Der Frameindex v,Ref beschreibt, dass zur Evaluation nur Frames $a_{v,\mathrm{Ref}}$ aus stimmhaften Regionen der Referenz herangezogen werden. Die Anzahl stimmhafter Frames wird mit $N_{v,\mathrm{Ref}}$ bezeichnet. Zwei wichtige Fehlergrenzen werden oft benutzt: das Fine-Error-Limit beschreibt eine Abweichunge $\Delta_{F_0}(a_{v,\mathrm{Ref}})$ von 5%

$$\Delta_{F_0,\mathrm{FE}}(a_{v,\mathrm{Ref}}) = 0,05 \cdot F_{0,\mathrm{Ref}}(a_{v,\mathrm{Ref}}) \tag{7.3}$$

und das Gross-Error-Limit beschreibt eine Abweichung von 20%

$$\Delta_{F_0,\mathrm{GE}}(a_{v,\mathrm{Ref}}) = 0,2 \cdot F_{0,\mathrm{Ref}}(a_{v,\mathrm{Ref}}). \tag{7.4}$$

Daraus lassen sich entsprechende Fehlerraten nach Tabelle 7.1 berechnen. In der Literatur sind die wichtigsten Maße die Gross Error Rate (GER) und die Fine Correct

Rate (FCR), die meist auch nur Correct Rate (CR, z. B. in [UH08c]) genannt wird. In Anlehnung an [YJM96] benutzt [KHK06] (und Andere) anstelle der GER noch die Unterteilung in untere und obere Gross Errors (GEL und GEH). Weitere existierende Fehlermaße sind die absolute Differenz der Mittelwerte aus Referenz und Schätzung sowie die absolute Differenz der Standardabweichungen aus Referenz und Schätzung [KHK06]. [UH08c] benutzt einen SNR, der die Leistung der Referenz $F^2_{0,\mathrm{Ref}}(a_{v,\mathrm{Ref}})$ zur Leistung des Fehlersignals $\Delta^2_{F_0}(a_{v,\mathrm{Ref}})$ ins Verhältnis setzt.

In der vorliegenden Arbeit wird vorwiegend die FCR benutzt. Dies hängt mit der Aufgabe des PDA, die er in HFA (vgl. Abschnitt 6) zu erfüllen hat, zusammen (grobe F_0-Schätzung ist nicht ausreichend). In einigen Fällen wird zusätzlich zur Verdeutlichung der Ergebnisse die GCR benutzt. Weitere Fehlermaße wurden teilweise gemessen, werden jedoch aus Gründen der Übersichtlichkeit nicht dargestellt.

7.3.4.2 VUD-Evaluationsparameter

Macht der VUD einen Fehler, handelt es sich um die folgenden zwei Möglichkeiten:

- **UE – Unvoiced Error**[7] beschreibt einen stimmlosen Frame $a_{u,\mathrm{Ref}}$, der fälschlicherweise als stimmhaft $a_{v,\mathrm{Est}}$ klassifiziert wurde, und
- **VE – Voiced Error**[8] beschreibt einen stimmhaften Frame $a_{v,\mathrm{Ref}}$, der fälschlicherweise als stimmlos $a_{u,\mathrm{Est}}$ klassifiziert wurde.

Aus beiden Fehlern lassen sich die entsprechenden, zur Evaluation herangezogenen Fehlerraten berechnen:

- Unvoiced Error Rate

$$\mathrm{UER} = \frac{N_{\mathrm{UE}}}{N_{u,\mathrm{Ref}}}, \tag{7.5}$$

  mit N_{UE} als Anzahl der UEs und $N_{u,\mathrm{Ref}}$ als Anzahl der stimmlosen Frames, und
- Voiced Error Rate

$$\mathrm{VER} = \frac{N_{\mathrm{VE}}}{N_{v,\mathrm{Ref}}}, \tag{7.6}$$

  mit N_{VE} als Anzahl der VEs und $N_{v,\mathrm{Ref}}$ als Anzahl der stimmhaften Frames.

Die Gesamtanzahl der Frames setzt sich aus

$$N_{\mathrm{ges}} = N_{u,\mathrm{Ref}} + N_{v,\mathrm{Ref}} \tag{7.7}$$

[7] Terminologie u. a. in [ECESS]. Hess [Hes08] benutzt den Term Voiced-to-Unvoiced Error.
[8] Terminologie u. a. in [ECESS]. Hess [Hes08] benutzt den Term Unvoiced-to-Voiced Error.

zusammen. Die Gesamtfehlerrate des VUDs, die hier mit ER (von engl.: Error Rate) bezeichnet ist, lässt sich nach

$$\mathrm{ER} = \frac{N_{\mathrm{UE}} + N_{\mathrm{VE}}}{N_{\mathrm{ges}}} \tag{7.8}$$

berechnen.

7.4 Basic-Extraktor-Verfahren

In diesem Abschnitt werden die Experimente zur Evaluation von PDAs beschrieben. Da, wie bereits erwähnt, nur die Basic-Extraktoren evaluiert werden, wird vereinfachend der Begriff BEA (engl.: Basic Extractor Algorithm) eingeführt, der hier auch für F_0-Detektionsmethode steht. Zunächst folgt ein kurzer Überblick über die verwendeten BEA-Methoden.

7.4.1 Überblick und Aufbau von BEA-Verfahren

Die Entwicklung und Untersuchung von BEAs reicht bereits weit in die Geschichte der Sprachsignalverarbeitung zurück. Die Anzahl von bislang entwickelten Ansätzen ist kaum überschaubar; jährlich werden neue Ansätze vorgestellt [QSS02]. Dabei kann zwischen traditionellen und neueren Verfahren unterschieden werden. Diese Arbeit gibt keinen Überblick zu Methoden der F_0-Detektion, dazu wird auf [Hes08] verwiesen, wo mehr als einhundert Ansätze zitiert werden. Im Folgenden sollen nur die getesteten Methoden kurz vorgestellt werden. Für genaue Informationen wird auf die entsprechenden Quellen verwiesen. Im Rahmen dieser Arbeit soll kein neuer Ansatz zur F_0-Detektion entwickelt werden, da der Fokus auf der Spracherkennung unter Hallbedingungen liegt. Anstelle dessen werden existierende Methoden auf ihre Funktionsfähigkeit unter Hallbedingungen getestet. Soweit es möglich ist, entspricht die Reihenfolge der zeitlichen Ordnung ihrer ersten Veröffentlichung. Die Bezeichnung der Methoden wird aufgrund der Einheitlichkeit und des Bezugs auf weitere Quellen in Englisch gehalten.

Da sich diese Evaluation an die Experimente von [UH08c, UHI08] anschließt, wurden zunächst die zwölf darin benutzten Methoden implementiert. Sie decken in etwa die wichtigsten derzeit verfügbaren BEA-Methoden ab. Weitere, z. B. zitiert in [Hes08], sind meist Abwandlungen eines Basisverfahrens [UH08c]. Zu diesen zwölf Methoden wurden hier noch zwei weitere (ACF und NCCF (s. u.)) hinzugefügt, sodass letztlich 14 Methoden evaluiert wurden.

Die Methoden haben Gemeinsamkeiten beginnend mit dem Framing, das, wie in Abschnitt 2.2.1 vorgestellt, eine Folge von Kurzzeitsignalabschnitten $\boldsymbol{x}(a)$ aus dem Signal ausschneidet (vgl. Gleichung (2.2)). Für die folgenden Erklärungen der einzelnen

Basic-Extraktor-Methoden soll nur auf die Analyse eines Frames eingegangen werden, weshalb hier aus Vereinfachungsgründen der Frameindex a weggelassen wird

$$x(a,k) = \boldsymbol{x} = x(k); \quad 0 \leq k \leq K-1. \tag{7.9}$$

Die einzelnen Basic-Extraktoren erzeugen für jeden Frame eine Analysefunktion $A(a,p)$, die hier vereinfacht mit $A(p)$ bezeichnet wird. In $A(p)$ wird dann nach dem Argument von Extremwerten gesucht

$$p_0 = \arg \max_{p_{0,\min}}^{p_{0,\max}} A(p) \quad \text{bzw.} \quad p_0 = \arg \min_{p_{0,\min}}^{p_{0,\max}} A(p), \tag{7.10}$$

die je nach Basic-Extraktor-Methode Maximum oder Minimum sein können. Die Variable p steht stellvertretend für eine Funktion der Frequenz f, die bspw. die Periodendauer $p = T = 1/f$, die Quefrenz $p = q = 1/f$, eine logarithmierte Frequenz nach (7.23) oder die Frequenz $p = f$ selbst sein kann. p_0 repräsentiert das entsprechende Pendant zur Grundfrequenz F_0, welche über die Umkehrung der im vorhergehenden Satz angesprochenen Funktion berechnet werden kann. Der Suchbereich wird zwischen $p_{0,\min}$ und $p_{0,\max}$ eingeschränkt, welche sich auf die Werte in (4.5) beziehen.

ACF – Auto Correlation Function Da es auf der Hand liegt, die Periodendauer eines periodischen Signals mit der Verschiebung des Maximums in der Autokorrelationsfunktion (ACF) zu messen, ist dies vermutlich die älteste Methode der PDA-Ansätze [Mar72]. Dabei wird nur der positive Zeitverschiebungsbereich $\kappa \geq 0$ der ACF $\psi_{xx}(\kappa)$ benötigt, der hier der Analysefunktion

$$A(\kappa) = \sum_{k=0}^{K-\kappa-1} x(k) \cdot x(k+\kappa) \tag{7.11}$$

entspricht. In $A(\kappa)$ wird nach (7.10) ein Maximum gesucht. κ entspricht der diskreten Zeitverschiebung, κ_0 der Grundperiodendauer T_0 beim Maximum von $A(\kappa)$. Es darf noch bemerkt werden, dass das Verhältnis $A(\kappa_0)/A(0)$ ein direktes Maß für die Deutlichkeit der Periodizität ist ($A(0)$ entspricht der Energie des Signalframes).

Cepstrum Die Cepstrum-Methode wurde 1964 von Noll et al. [Nol64, Nol66] vorgestellt. Aufgrund der Nichtsinusform des Anregungssignales finden sich im Amplitudenspektrum $|\underline{X}(n)|$ Oberschwingungen zu Vielfachen der Grundfrequenz, die Harmonischen. Bildet man das logarithmierte Amplitudenspektrum $\lg(|\underline{X}(n)|)$, verschwindet seine hohe Dynamik und die harmonische Struktur ist in $\lg(|\underline{X}(n)|)$ als periodische Funktion (der Frequenz) mit der Periode n_0 (zugehörig zu F_0) sichtbar. Durch Fouriertransformation

$$A(q) = \mathrm{DFT}\left\{\lg|\underline{X}(n)|\right\} \tag{7.12}$$

erhält man als Analysefunktion das Cepstrum, in dem sich bei der zur Grundfreqzenz zugehörigen Quefrenz q_0 ein Maximum abzeichnet. Dieses wird wieder nach (7.10) gesucht. Man kann diese Operation auch abstrakt als Fourieranalyse einer periodischen Funktion interpretieren, in deren Ergebnis sich an der (abstrakten) Grundfrequenz ein Maximum befindet. Aus dieser gedachten Grundfrequenz kann durch die Bildung des Reziproken auf die abstrakte Grundperiode, die ja die eigentliche gesuchte Grundfrequenz F_0 ist, geschlossen werden.

AMDF – Average Magnitude Difference Function Die Methode wurde 1968 in [SB68] (Quellenangabe nach [Hes08]) bzw. 1974 in [RSC+74] vorgestellt. Sie bildet die Summe der Differenzfunktion zwischen zwei um κ verschobenen Funktionen

$$A(\kappa) = \sum_{k=0}^{K-1} |x(k) - x(k+\kappa)| \,. \tag{7.13}$$

Bei idealer Periodizität verschwindet bei der Verschiebung um die Periodendauer κ_0 die Analysefunktion $A(\kappa)$, (7.10) wird deshalb zu einer Minimumsuche. κ entspricht wieder der diskreten Zeitverschiebung, κ_0 der Grundperiodendauer T_0 beim Maximum von $A(\kappa)$.

ACLS – Auto Correlation of Logarithmized Amplitude Spectrum Ähnlich wie die Cepstrum-Methode sucht ACLS [KSS97] nach einer Periodizität in $\lg(|\underline{X}(n)|)$. Die Periodizität wird durch eine Autokorrelationsfunktion im Frequenzbereich

$$A(\nu) = \sum_{n=0}^{N-\nu-1} \lg(|\underline{X}(n)|) \cdot \lg(|\underline{X}(n+\nu)|) \tag{7.14}$$

ermittelt. Mit (7.10) wird nach dem Maximum gesucht, welches sich bei der zur Grundfrequenz identischen Frequenzverschiebung ν_0 einstellt.

LP Residual Die Lineare Prädiktion zur Analyse eines Zeitsignalabschnittes $x(k)$ versucht, den Vokaltrakt $v_{\mathrm{Ph}}(k)$ (vgl. Gleichung (4.1)) eines Menschen durch ein Prädiktionsfilter nachzubilden. Bei optimaler Filterschätzung und anschließender Filterung von $x(t)$ mit dem inversen Prädiktorfilter (inverser Vokaltrakt) entsteht der Prädiktionsfehler (engl.: LP Residual), der als geschätztes Anregungssignal $e'_{\mathrm{Ph}}(k)$ interpretiert werden kann [VHH98]. In $e'_{\mathrm{Ph}}(k)$ tritt F_0 bzw. T_0 deutlicher hervor. Die LP-Residual-Methode zur F_0-Detektion versucht nun, F_0 aus dem geschätzten Anregungssignal zu ermitteln. Dazu kann ein beliebiger Basic-Extraktor benutzt werden (z. B. ACF, AMDF, Cepstrum etc.). Markel stellt die Methode 1972 als erster unter dem Namen Simplified Inverse Filter Tracking (SIFT) vor [Mar72]. Zur F_0-Extraktion benutzt er eine ACF auf $e'_{\mathrm{Ph}}(k)$. Hinzu kommt noch eine Tiefpassfilterung,

um F_0 zu betonen. Die LP-Residual-Methode stellt eigentlich nur eine robustheitssteigernde Komponente zu Basisverfahren wie der ACF dar. Das LP-Filter hat ein zum Sprachspektrum inverses Spektrum. Demnach hat das generierte Residual-Signal ein weißes Spektrum. Daraus resultieren zwei Vorteile. Zum einen wird die F_0-Detektion robuster gegen Fehlschätzungen hin zu besonders betonten Formanten (typisches F_0-Detektionsproblem). Zum anderen werden spektral basierte Methoden (Cepstrum etc.) robuster, da die Periodizität im Spektrum nun relativ konstant über das gesamte Spektrum hervortritt (weißes Spektrum, vgl. auch Graphik in auf Seite 183 in [VHH98]). Für die Evaluation wurde eine LP-Filterordnung von 12 gewählt. F_0 wurde im Anschluss mit der ACF-Methode auf dem LP-Residual-Signal extrahiert.

ACLag – Auto Correlation of Lagwindowing on Amplitude Spectrum Sagayama stellt 1978 [Sag78] (zitiert in [SS92]) die Methode ACLag vor (Blockdiagramm in [SS92]). Sie berechnet zunächst das Leistungsspektrum $P(f) = |\underline{X}(f)|^2$. Mit der IFFT wird aus dem Leistungsspektrum die Autokorrelationsfunktion berechnet. In der ACF wird der Lag-Bereich ausgeblendet (Lag-Windowing). Dies entspricht dem Pendant zur cepstralen Glättung im Cepstrum. Die nachfolgende FFT führt zu einem geglätteten Leistungsspektrum $P_s(f)$ (Eliminierung der harmonischen Struktur). Die Division

$$E'^2_{\mathrm{Ph}}(f) = \frac{P(f)}{P_s(f)} \tag{7.15}$$

normiert (weißt) das Leistungsspektrum, was zur geschätzten spektralen Struktur des Anregungssignales E'^2_{Ph} führt. Die Operation kann auch als Umkehrung der quadrierten Gleichung (4.2) betrachtet werden; $P_s(f)$ wird dann als V^2_{Ph} interpretiert. Nach IFFT bildet sich die ACF von e'_{Ph}, welche ein starkes Maximum bei T_0 aufweist. Ähnlich wie die LP-Residual-Methode kann die ACLag-Methode als robustheitssteigerndes Element, gefolgt von einem Basisverfahren, gesehen werden.

ACMWL – Auto Correlation through Multiple Window Length Wie der Name bereits andeutet, basiert die Methode ACMWL nach [TSM79] auf der ACF-Methode. Jedoch wird hier die ACF nach (7.11) für einen einzelnen Frame über eine sukzessiv steigende Framelänge N ($1 \leq N \leq N_{\max}$) durchgeführt (zusätzliche Normierung auf die aktuelle Framelänge). Für jede Framelänge wird eine Maximumsuche in der ACF nach (7.10) durchgeführt und die gefundenen Maxima $\hat{\psi}_{xx}$ sowie ihre zugehörigen Verschiebungen κ'_0 bilden je einen Vektor mit $N_{\max}$ Elementen $\hat{\psi}_{xx}(N)$ und $\kappa'_0(N)$. Die Entscheidungsfunktion $A(N)$ ist eine Funktion der variierten Framelänge N und wird aus einer akkumulierten Stärke der gefundenen Maxima ermittelt, die mit der Wichtungsfunktion $g_N(n)$ bewertet werden

$$A(N) = \hat{\psi}_{xx}(N) + \sum_{n=1}^{N_{\max}} \hat{\psi}_{xx}(n) \cdot g_N(n). \tag{7.16}$$

Die Stärke der Maxima liefert ja bereits die Aussage, wie deutlich die Periodizität im Signal zu erkennen ist. $g_N(n)$ ist eine Ausblendefunktion für besonders stark abweichende κ_0'. Für ein bestimmtes N wird die Funktion der Abweichung für jedes n aus dem Verhältnis

$$d_N(n) = \left| \frac{\kappa_0'(N)}{\kappa_0'(n)} - 1 \right|; \quad 1 \leq n \leq N_{\max} \tag{7.17}$$

berechnet. $d_N(n)$ schwankt für geringe Abweichungen um 0. Größere Abweichungen sollen in (7.16) gedämpft bzw. ausgeblendet werden. Deshalb folgt

$$g_N(n) = \begin{cases} 1 & ; \quad d_N(n) \leq 0,1 \\ (0,25 - d_N(n)) \cdot 0,15 & ; \quad 0,1 \leq d_N(n) \leq 0,25 \\ 0 & ; \quad 0,25 \leq d_N(n). \end{cases} \tag{7.18}$$

Anschließend wird das Maximum in $A(N)$ gesucht, das sich bei N' befindet. F_0 korrespondiert letztlich zu der Verschiebung $\kappa_0'(N')$.

Im Vergleich zur ACF ergibt sich bei der ACMWL der Vorteil, dass das κ_0 bei der zeitlich optimalen Fensterlänge ermittelt wird.

Comb – Kammfilter im Frequenzbereich In [Mar82] wird die Kammfilter-Methode vorgestellt. Sie arbeitet ebenfalls auf $\lg(|\underline{X}(n)|)$ und versucht, die durch die Harmonischen beschriebene periodische Funktion im Spektrum zu analysieren. Das Kammfilter besteht aus einer Überlagerung von Bandpassfiltern $G_{\mathrm{BP},f_{\mathrm{BP}}}(n)$, deren Mittenfrequenzen f_{BP} Vielfache einer gedachten Grundfrequenz F_0' sind. Es entsteht ein kammartiger Frequenzgang

$$G_{\mathrm{Comb}}(F_0', n) = \sum_{i=1}^{N_{\mathrm{Comb}}} W(n) \cdot G_{\mathrm{BP},i \cdot F_0'}(n), \tag{7.19}$$

der noch mit einer abfallenden Wichtungsfunktion $W(n)$ multipliziert wird. Dieses Filter wird für alle zu testenden F_0 erzeugt; es entsteht eine Art Kammfilterbank, deren einzelne Filterausgänge je einer zu testenden F_0 zugeordnet werden. Das logarithmierte Amplitudenspektrum wird nun mit diesen Filtern im Frequenzbereich multipliziert und danach werden die Integrale (Summen) über die Filterausgangsspektren gebildet (ähnlich wie die Energie). Daraus entsteht eine von F_0' abhängige Analysefunktion

$$A(F_0') = \sum_{n=0}^{N-1} \left\{ G_{\mathrm{Comb}}(F_0', n) \cdot \lg |\underline{X}(n)| \right\}, \tag{7.20}$$

in der sich bei der realen F_0 ein Maximum einstellt, welches erneut durch (7.10) ermittelt wird. Man kann bei der Kammfiltermethode noch Spielarten variieren, bspw. die Benutzung des Leistungsspektrums anstelle des Amplitudenspektrums etc.

SHS – Subharmonic Summation Hermes stellt 1988 in [Her88] die Methode der Subharmonic Summation vor. Dabei wird als Ausgangspunkt ein Leistungsdichtespektrum $S(f)$ gewichtet

$$P(f) = W(f) \cdot S(f). \tag{7.21}$$

Die Wichtungsfunktion $W(f)$ wird durch eine Arcustangensfunktion gebildet, die sehr tiefe Frequenzen ausblendet. Die Frequenzachse von $P(f)$ wird durch einen sich erhöhenden Faktor i gestaucht ($P(i \cdot f)$). Die gestauchten Spektren werden aufaddiert

$$A(f) = \sum_{i=1}^{N} h_i \cdot P(i \cdot f). \tag{7.22}$$

Bei der Grundfrequenz F_0 ergibt sich in $A(f)$ ein Maximum durch Überlagerung der Spektralkomponenten an den Vielfachen von F_0. [Her88] empfiehlt weiterhin eine logarithmierte Frequenzachse

$$s = \mathrm{ld} f, \tag{7.23}$$

wobei die Stauchung nun in eine Verschiebung nach links resultiert. Daraus ergibt sich dann die für SHS typische Analysefunktion

$$A(s) = \sum_{i=1}^{N} h_i \cdot P(s + \mathrm{ld} i). \tag{7.24}$$

Das Maximum in $A(s)$ tritt hier noch deutlicher hervor als in (7.22), da durch die Logarithmierung Nebenmaxima besser gestreut werden (graphische Erläuterung in [Her88]). Der Wichtungsfaktor h_i ist nach

$$h_i = 0,84^{i-1}; \quad 1 \leq i \leq N \tag{7.25}$$

definiert und stellt eine fallende Exponentialfunktion dar, die stärkere Stauchungen bzw. Verschiebungen geringer bewertet. N wird in [Her88] mit 15 vorgeschlagen.

TEMPO – Instantanious Frequency Für einige Problemstellungen reicht die zeitliche Auflösung der frameweisen Abarbeitung nicht aus. Charpentier stellt 1986 in [Cha86] das Prinzip der Momentanfrequenz ω_M (Kreis- oder Winkelfrequenz, engl.: Instantanious Frequency) zur Bestimming von F_0 vor. Dazu wird das Signal zunächst mit einem Bandpass $g_{\mathrm{BP},F_0'}(n)$ um eine angenommene Grundfrequenz F_0' gefiltert, sodass das Filterausgangssignal $x_{F_0'}(t)$ entsteht. Die Momentanfrequenz ist über die

Ableitung der Phase eines analytischen Signals $\underline{x}_{a,F_0'}(t)$, welches sich mit der Hilberttransformation $\mathcal{H}\{\cdot\}$ berechnen lässt, definiert

$$\underline{x}_{a,F_0'}(t) = x_{F_0'}(t) + j\mathcal{H}\left\{x_{F_0'}(t)\right\}, \tag{7.26}$$

$$\underline{x}_{a,F_0'}(t) = A_{a,F_0'}(t) \cdot e^{j\varphi_{a,F_0'}(t)}, \tag{7.27}$$

$$\omega_M(F_0', t) = \frac{\mathrm{d}\varphi_{a,F_0'}(t)}{\mathrm{d}t}. \tag{7.28}$$

D. h., für jeden Zeitpunkt t lässt sich eine Funktion $\omega_M(F_0')|_t$ bestimmen. Sie hat die Eigenschaft, bei Variation von F_0' in der Nähe von F_0 bzw. bei deren Vielfachen (Harmonischen) zu einer konstanten Funktion zu werden mit

$$F_0' \approx F_0 \quad \rightarrow \quad \omega_M(i \cdot F_0')|_t = i \cdot F_0; \qquad i \in \mathbb{N}, i \neq 0. \tag{7.29}$$

Diese Plateaus (graphische Erläuterung in [Cha86, KKdCP99, AIK+00]) können durch verschiedene Verfahren gefunden werden (z. B. Histogrammansatz etc.), die als Ergebnis die momentane F_0 liefern. Es existieren verschiedene Ansätze für die Implementation der Methode der Momentanfrequenz. Im Rahmen dieser Versuchsreihe wurde das Verfahren nach Kawahara [KKdCP99] benutzt, das auch als TEMPO bekannt ist. Es sucht das Plateau der Grundfrequenz, welches sich durch ein großes Träger-zu-Rausch Verhältnis am deutlichsten abzeichnet (genauere Erläuterungen in [KKdCP99]).

NCCF – Normalized Cross Correlation Function Die Methode ist eine normalisierte Spielart der ACF und wurde 1995 in [Tal95] vorgestellt. Die NCCF berechnet sich nach

$$A(\kappa) = \frac{\sum_{k=0}^{K-1} x(k) \cdot x(k+\kappa)}{\sqrt{\left[\sum_{k=0}^{K-1} x^2(k)\right]\left[\sum_{k=0}^{K-1} x^2(k+\kappa)\right]}}, \tag{7.30}$$

wobei auch hier wie bei der ACF-Methode mittels (7.10) der Index κ_0 des Maximums von $A(\kappa)$ gesucht wird. In [Tal95] wird berichtet, dass die Methode stabiler arbeitet als die reine ACF, da die Maxima besser hervortreten und weniger durch schnelle Wechsel der Signalamplitude beeinflusst werden.

IFHC – Instantanious Frequency of Harmonic Components Die IHCF ist eine Weiterentwicklung von TEMPO und wurde 2000 von Atake et al. [AIK+00] vorgestellt. Sie soll besonders die Geräuschrobustheit von TEMPO steigern. Das Grundprinzip arbeitet wie bei TEMPO. Allerdings werden im Unterschied dazu nicht nur die F_0-Komponente, sondern zusätzlich verschiedene Harmonische von F_0 mit in die Messung

einbezogen. TEMPO wird zunächst zur Grobschätzung von F_0 benutzt. Die Filterung geschieht erneut, danach aber im Frequenzbereich; zuvor werden die Fensterfunktionen für die FFT auf die grob geschätzte F_0 optimiert (Die Begründung ist ähnlich wie bei ACMWL.). Mit dieser optimierten Funktion $\omega_{M,opt}(i \cdot F_0')|_t$ werden die Plateaus der F_0 und der harmonischen Komponenten bestimmt, die letztlich noch mit einer adaptiven Wichtungsfunktion beaufschlagt werden. Für die genaue Implementierung von IFHC wird auf [AIK+00] verwiesen.

PHIA – Periodicity/Harmonicity using Instantanious Amplitudes PHIA wird 2001 von Ishimoto et al. in [IUA01] vorgestellt. Dabei wird F_0 basierend auf einer Messung der Periodizität und Harmonizität der Momentanamplitude geschätzt. Die Periodizität wird mittels einer Filterbank mit konstanter relativer Bandbreite (256 Kanal-Gammaton-Filterbank) gemessen und die Harmonizität wird mittels einer Filterbank mit konstanter absoluter Bandbreite (256 Kanal-Gammaton-Filterbank oder FFT als Filterbank) gemessen. Zur Funktionsweise vgl. Graphik in [IUA01]. Für die Periodizität werden aus den Filterbankausgangssignalen F_0-Kandidaten mit der ACF im Zeitbereich ermittelt. Gleiches geschieht für die Harmonizität mittels ACF im Frequenzbereich. Für beide Kandidatenmengen (F_0-Kandidaten für Periodizität bzw. für Harmonizität) werden mit einem Histogrammansatz die Wahrscheinlichkeiten der F_0 ermittelt. Aus beiden Wahrscheinlichkeitsfunktionen wird F_0 geschätzt. [IUA01] berichtet, dass PHIA besonders bei starken Geräuschen robuster als andere Algorithmen arbeitet (konkret: Vergleich mit Cepstrum und TEMPO). Zusätzlich stellt [IUA01] ein Verfahren vor, welches PHIA nur als Grobschätzung benutzt. Die grob geschätzte F_0 dient als Parameter für einen Geräuschunterdrückungsschritt basierend auf einer adaptiven Kammfiltermethode (abhängig von F_0). Letztlich wird TEMPO zur F_0-Detektion auf das geräuschreduzierte Signal verwendet.

YIN Der YIN-Schätzer wurde von de Cheveigné et al. 2002 vorgestellt. Er stellt eine Art Kombination von ACF und AMDF dar, wobei die AMDF nicht in Reinform benutzt wird, sondern anstelle der Beträge in Gleichung (7.13) werden die Quadrate der Differenzen aufsummiert. Zusätzlich wird diese abgewandelte AMDF nun noch auf ihre Energie in den unteren Verschiebungen normiert. Aus diesen Bausteinen entsteht die so bezeichnete „Cummulative Mean Normalized Difference Function“, die als Analysefunktion $A(p)$ dient. Insgesamt werden sechs Schritte benötigt, um das Ergebnis zu berechnen; für die ausführliche Beschreibung des YIN-Schätzers wird auf [dCK02] verwiesen.

CCA – Complex Cepstrum Analysis Die CCA, vorgeschlagen 2008 von Unoki et al. [UH08c, UHI08], ist die erste und derzeit einzige F_0-Detektionsmethode, die versucht, Robustheit gegen Raumhallstörungen zu erreichen. Dies geschieht im Wesentlichen in einer Art „Enthallungsschritt“, der jedoch nicht die RIR eliminiert, sondern Teile

ihres komplexen Cepstrums. Dazu wird zunächst vorausgesetzt, dass das Spektrum eines Signals bzw. der Frequenzgang eines Systems in seine Mindestphasen- und Allpasskomponenten zerlegt werden kann, welche nach Betrag und Phase geordnet notiert werden

$$\underline{H}(\omega) = |\underline{H}_{\min}(\omega)| \cdot e^{j \arg\{\underline{H}_{\min}(\omega)\}} \cdot |\underline{H}_{\text{all}}(\omega)| \cdot e^{j \arg\{\underline{H}_{\text{all}}(\omega)\}}. \tag{7.31}$$

Im zugehörigen Cepstrum ergibt sich die additive Dekomposition

$$c(q) = c_{\text{A},\min}(q) + c_{\phi,\min}(q) + c_{\text{A,all}}(q) + c_{\phi,\text{all}}(q), \tag{7.32}$$

wobei die Indexnotationen A auf das Amplitudenspektrum und ϕ auf das Phasenspektrum verweisen. Außerdem entfällt der Term $c_{\text{A,all}}(q)$, da der Logarithmus von 1 (Frequenzgang eines Allpasssystems) gerade 0 ist. Betrachtet man ein verhalltes Sprachsignal (Frequenzbereich)

$$X(\omega) = S(\omega) \cdot H(\omega), \tag{7.33}$$

so ergibt sich die entsprechende Dekomposition im Cepstralbereich nach

$$\begin{aligned} c(q) \quad &= \quad c_{S,\text{A},\min}(q) + c_{S,\phi,\min}(q) + c_{S,\phi,\text{all}}(q) \\ &+ \quad c_{H,\text{A},\min}(q) + c_{H,\phi,\min}(q) + c_{H,\phi,\text{all}}(q). \end{aligned} \tag{7.34}$$

[UH08c] stützt seine Annahmen auf Experimente, bei denen künstlich einmal $c_{H,\text{all}}(q)$ bzw. $c_{H,\min}(q)$ aus $c(q)$ eliminiert wurde. Nachfolgende Signalrekonstruktion und F_0-Detektion ergaben, dass vorwiegend die Allpasskomponente der RIR $c_{H,\text{all}}(q)$ einen störenden Einfluss auf die Grundfrequenz hat. Somit versucht CCA, genau diese Komponente, die sich in (7.33) bereits auf $c_{H,\phi,\text{all}}(q)$ beschränkt, zu ermitteln und zu eliminieren. Dafür wird zunächst die Nachhallzeit blind aus $x(t)$ mit einer MTF-basierten Methode geschätzt, die, ebenfalls von Unoki et al., bereits bei einem MTF-basiertem Enthallungsalgorithmus vorgestellt wurde [USFA04, UFSA04]. Mit der geschätzten Nachhallzeit und der Formel von Schröder (3.81) wird ein künstliches $\hat{h}(t)$ für die Berechnung von $c_{H,\text{all}}(q)$ erzeugt. An dieser Stelle erkennt man zwei mögliche Schwächen des Verfahrens: a) es ist nur für das Fernfeld ausgelegt, da sowohl die MTF-basierte Schätzung als auch $\hat{h}(t)$ auf einer Fernfeldannahme beruhen; b) $\hat{h}(t)$ ist kein genaues Abbild von $h(t)$, da die Schätzung von T_{60} ungenau sein kann. Nach Subtraktion von $c_{\hat{H},\phi,\text{all}}(q)$ wird das Signal wieder in den Zeitbereich überführt

$$\hat{x}(t) = \text{IFFT}\left\{ e^{\text{FFT}\{c(q) - c_{\hat{H},\phi,\text{all}}(q)\}} \right\}. \tag{7.35}$$

Es folgt eine BEA-Methode, die nun wieder framebasiert auf $\hat{x}(t)$ arbeitet. Dabei werden die Schritte (1) Cepstrum, (2) Langpasslifter, um das Anregungssignal zu erhalten, (3) FFT, um in den logarithmierten Spektralbereich zurück zu gelangen, sowie

(4) Comb (Kammfiltermethode), um letztlich F_0 zu ermitteln, abgearbeitet. Die Extraktion aus $\hat{x}(t)$ wäre auch mit einer anderen Methode, wie z. B. der ACF, möglich. Die Tatsache, dass $\hat{h}(t)$ nur eine statistische Schätzung von $h(t)$ ist, da das exponentiell bewertete Geräusch in (3.81) eine rein zufällige Größe ist und damit der originale Verlauf von $h(t)$ keineswegs nachgebildet wird, ist keine Fehlerquelle. Der Grund dafür liegt in den Eigenschaften von Allpass und Mindestphasensystem. Tatsächlich wird ja nur $c_{\hat{H},\phi,\mathrm{all}}(q)$ subtrahiert, welches eine Phasenmanipulation darstellt. Diese hat auf eine nachfolgende F_0-Extraktionsstufe kaum Einfluss, wenn diese bspw. auf dem Amplituden- oder Leistungsspektrum arbeitet. Trotz der angesprochenen Schwächen zeigen [UH08c] sowie auch die folgenden Experimente, dass die Methode zur Verbesserung der F_0-Detektion unter Hallbedingungen (nur Fernfeld) beitragen kann.

7.4.2 Evaluation von Basic-Extraktor-Verfahren unter Hallbedingungen

Die Experimente werden getrennt für männliche und weibliche Sprache durchgeführt. Sie unterteilen sich in die vier Hallabhängigkeiten (a) bis (d), wie sie in Abschnitt 7.3.3 vorgestellt wurden. Aus Gründen der Übersichtlichkeit sind in den Diagrammen nur die Ergebnisse der sieben wichtigsten Methoden dargestellt. Für die Auswahl waren dabei zwei Kriterien interessant:

- Methoden, die in der Evaluation am besten abgeschnitten haben (ACF, Cepstrum, SHS, CCA, NCCF), und
- Methoden, die besonders genau arbeiten (TEMPO, YIN)[9] und deshalb in vielen Evaluationen zum Vergleich herangezogen werden [UH08c].

Die anderen sieben Methoden schnitten in der Evaluation teilweise schlechter, teilweise ähnlich gut ab. Die Darstellung ihrer Ergebnisse ist deshalb hier nicht relevant; es darf aber ausgesagt werden, dass sie nicht besser abschneiden als die besten dargestellten Methoden. Einige ihrer Ergebnisse sind im Vergleich in [UH08c] unter künstlichen Hallbedingungen dargestellt.

7.4.2.1 Abhängigkeit von T_{60} bei Verhallung mit künstlich generierten RIRs

Die erste Abhängigkeit nach Abschnitt 7.3.3 ist (a) das Verhalten der FCR als Funktion der Nachhallzeit bei Verhallung mit künstlich generierten RIRs. Sie werden nach dem Modell von RIRs (3.81) generiert. Die Direktschallphase wird nicht hinzugezogen, somit beschreiben die künstlichen RIRs das Fernfeld in Räumen mit den entsprechenden Nachhallzeiten. Für diese Bedingungen ist diese Evaluation zulässig. Es entsteht

[9] Beide Methoden arbeiten sehr genau in sauberen Umgebungsbedingungen, sind jedoch nicht robust gegen Rauschen [UH08c]. Wegen seiner Genauigkeit wird TEMPO auch manchmal zur Erstellung von $F_{0,\mathrm{Ref}}$ auf ein EGG-Signal angewendet [UH08c].

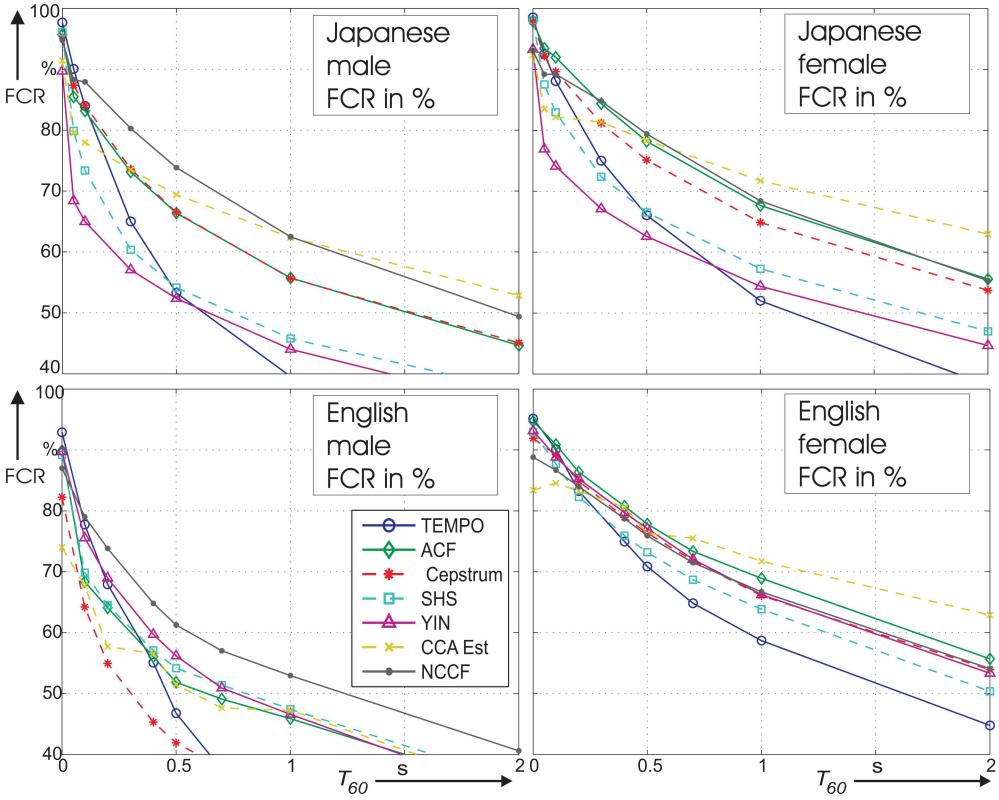

Abbildung 7.2 – FCR für die Verhallung mit künstlichen RIRs. Vergleich mit den Ergebnissen der Studie aus [UH08c] mit der japanischen Datenbasis (oben) und der kleinen englischen Datenbasis (unten). Links: männliche Sprecher, rechts: weibliche Sprecher.

noch der Unterschied zu realen RIRs, dass eine Frequenzabhängigkeit des Halls nicht vorhanden ist. Dies entspricht zwar nicht der Realität, ist jedoch in einem wichtigen Punkt ein Vorteil: Es besteht die Möglichkeit, unterschiedlichste Datenbasen definiert zu verhallen, ohne dass man auf real gemessene RIRs zurückgreifen muss. Reale RIRs variieren selbst bei gleichen Nachhallzeiten von Raum zu Raum; künstliche RIRs hingegen können mathematisch genau ausgedrückt werden. Dadurch werden Messungen verschiedener Experimente zu Hallabhängigkeiten mit verschiedenen Datenbasen miteinander vergleichbar. Somit ist es auch möglich, an die Experimente in [UH08c] anzuschließen, indem zunächst ein Vergleich der Methoden über die japanische Datenbasis **(i)** (entspricht den Ergebnissen in [UH08c]) und die kleine englische Datenbasis **(ii)** durchgeführt wird.

Die Ergebnisse in Abbildung 7.2 zeigen, dass die FCR erwartungsgemäß für alle Me-

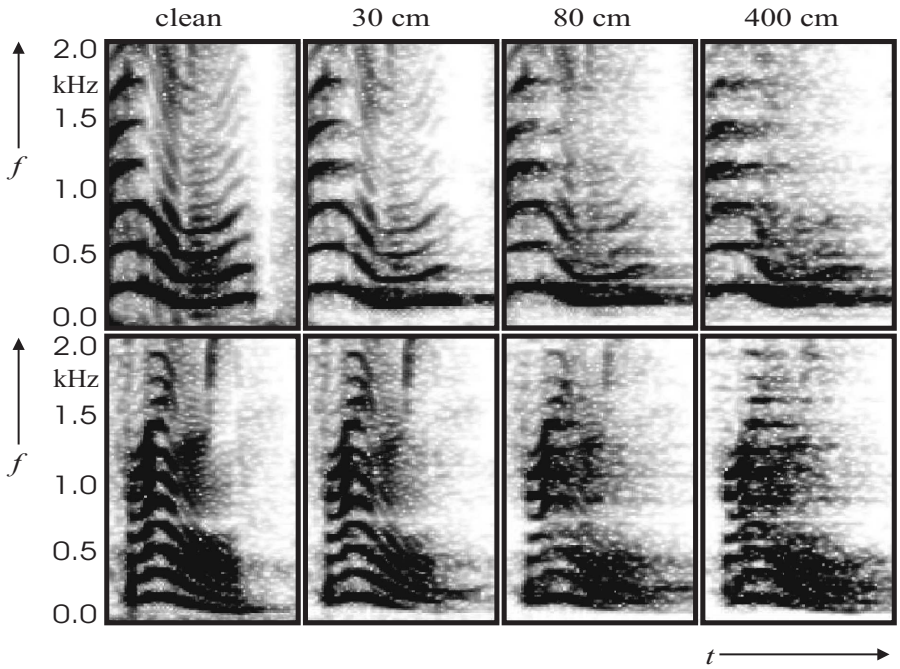

Abbildung 7.3 – Spektrogramme von 0,8 s verhallter Sprache bei steigendem SMD (steigende Verhallung; unverhallt, SMD = 30 cm, SMD = 80 cm und SMD = 400 cm) im SMART-Room. Oben: weiblicher Sprecher, unten: männlicher Sprecher.

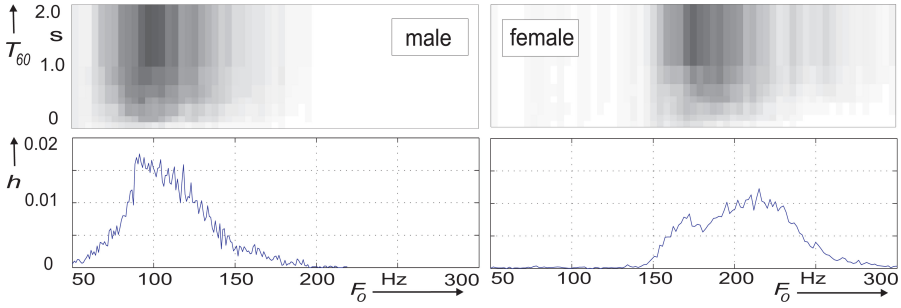

Abbildung 7.4 – Oben: Anzahl der Fehler (Fine Errors) in Abhängigkeit von der Nachhallzeit (Ordinate) bezogen auf ein F_0-Frequenzband (Abszisse). Für die Darstellung sind die Ergebnisse des YIN benutzt worden, andere BEAs zeigen ähnliches Verhalten. Unten: relative Häufigkeit der Grundfrequenz F_0 in der Datenbasis, gewonnen aus den Referenzdaten. Alle vier Graphiken sind unter Benutzung der kleinen englischen Datenbasis erstellt worden. Links: männliche Sprecher, rechts: weibliche Sprecher.

thoden mit steigender T_{60} abfällt (eine Bewertung (Reihenfolge) der Methoden soll an dieser Stelle nicht gegeben werden). Für beide Datenbasen ist erkennbar, dass die FCR für weibliche Sprecher deutlich höher ist als für männliche. Zusätzlich funktionieren die BEAs generell besser für die japanische Datenbasis im Vergleich zur englischen. Um diese Unterschiede zu erklären, wurde eine etwas genauere Fehleranalyse durchgeführt, welche in Abbildung 7.4 (für die englische Datenbasis) dargestellt ist. Betrachtet man die Fehlerverteilung (oben) im Vergleich zur F_0-Verteilung (unten), ist zu erkennen, dass das Ansteigen der Fehler mit steigender T_{60} für männliche Sprecher etwa der Verteilung der F_0 entspricht. Für weibliche Sprecher stellt sich jedoch bei steigender T_{60} ein deutlich bemerkbarer Überhang an Fehlern ein, der zu niedrigen F_0 tendiert. Dieser Effekt wird erklärbar, wenn man sich die Spektrogramme verhallter Sprachdaten in Abbildung 7.3 ansieht. Man erkennt den Unterschied in der Grundfrequenz und in den Harmonischen. Bei der tieferen männlichen Grundfrequenz liegen die Harmonischen dichter beieinander als bei der höheren weiblichen Grundfrequenz. Bei Verhallung zieht ein Schallereignis eine Hallschleppe hinter sich her, die eine Art bandbegrenztes Rauschen bei der Ursprungsfrequenz des Soundereignisses darstellt. Man sieht dies in der Abbildung für die Harmonischen. Variiert die Grundfrequenz, wie dargestellt, dann bewegen sich die Harmonischen (je nach Richtungsänderung nach oben oder unten) in die Hallschleppe der entsprechenden Nachbarharmonischen hinein. Die nun hineinragende Hallschleppe repräsentiert eine Störung. Variiert F_0 nicht ganz so stark, ragt die Hallschleppe in den Zwischenraum zweier Harmonischer hinein und stört somit die harmonische Struktur (Periodizität in Frequenzrichtung) des Spektrums. Beide Effekte wirken sich negativ auf das Verhalten der BEAs aus. Bei geringerer Grundfrequenz wirken die beiden Effekte stärker als bei höherer Grundfrequenz, wo der Abstand zwischen zwei Harmonischen größer ist. In Abbildung 7.3 ist bereits durch subjektive Betrachtung zu erkennen, dass die harmonische Struktur für männliche Sprache stärker gestört ist als für weibliche. Bezieht man noch die Werte der mittleren Grundfrequenz aus Tabelle 4.1 in die Überlegungen ein, so lässt sich auch erklären, warum die BEAs für die japanische Datenbasis bessere Ergebnisse liefern. Dies unterstreicht ebenfalls die vorgeschlagene Erklärung zur Abhängigkeit der BEAs vom Geschlecht der Sprecher.

7.4.2.2 Abhängigkeit vom SMD bei konstanter T_{60}

Im Unterschied zur künstlichen Fernfeldsimulation im vorigen Abschnitt wird nun das Verhalten der BEAs bei Variation des SMD getestet. Zur Verhallung wurden die gemessenen RIRs im SMART-Room benutzt. Die Nachhallzeit ist demnach fixiert bei ca. 0,7 s. Die Ergebnisse in Abbildung 7.5 (oben) zeigen, dass die FCR für alle Methoden mit zunehmender Verhallung abfällt. Ein großer Unterschied entsteht zwischen den FCRs bei weiblichen und männlichen Sprechern. Für männliche Sprecher sinkt die F_0-Detektionsleistung bereits drastisch nach einem SMD von wenigen Zentimetern. Für weibliche Sprecher hingegen ist das Sinken der FCR moderat und liegt noch im für

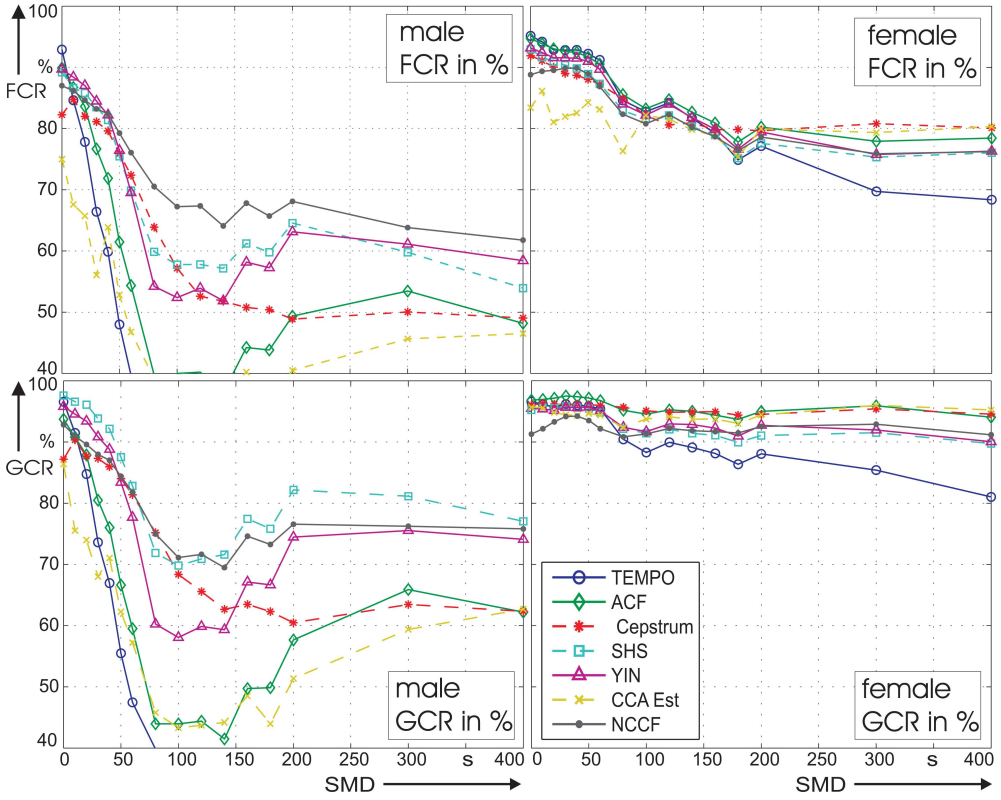

Abbildung 7.5 – Abhängigkeit vom SMD in der SMART-Room Umgebung. Oben: FCR; unten: GCR (zur Verdeutlichung der Ergebnisse). Links: männliche Sprecher, rechts: weibliche Sprecher.

die Zielanwendung (vgl. HFA) akzeptablen Bereich.

Um den Unterschied der F_0-Detektionsleistung bei männlicher und weiblicher Sprache zu verdeutlichen, wird hier ausnahmsweise zusätzlich die GCR benutzt (Abbildung 7.5 (unten)). Die GCR hat mit dem Gross-Error-Limit von 20 % einen sehr großen Toleranzbereich für Fehler. Die Graphik stellt somit dar, dass die F_0-Detektionsleistung für männliche Sprache selbst bei größerer Toleranz noch immer stark einbricht und somit die Masse der Fehlschätzungen der FCR außerhalb der 20-%-Grenze liegt. Bei den weiblichen Sprechern ist genau das Gegenteil zu erkennen. Die GCR hat für mehrere Methoden einen fast ungestörten bzw. gering gestörten Verlauf. Daraus ist zu schließen, dass die Masse der Fehler innerhalb der 20-%-Grenze liegt. Die Zusammenstellung der Diagramme in Abbildung 7.5 stellt somit besonders deutlich den Unterschied der F_0-Detektionsleistung von Sprachdaten, die von weiblichen bzw. männlichen Spre-

chern stammen, heraus. Besonders bemerkenswert ist der steile Abfall bei männlichen Sprechern nach einem SMD von bereits wenigen Zentimetern.

Um einen Bezug zwischen den Experimenten in Abbildung 7.2 und 7.5 herzustellen, kann man die FCRs in Abblidung 7.5 (oben) bei SMD = 400 cm (angenommenes Fernfeld) ablesen und diese mit den FCRs in Abblidung 7.2 (oben, englische Datenbasis) bei T_{60} = 0,7 s (entspricht etwa dem SMART-Room) vergleichen. Die Bereiche der abgelesenen FCRs stimmen in etwa überein:

Abbildung	Bedingung		männlich	weiblich
Abbildung: 7.2	$T_{60} = 0,7\ s$	SMD = ∞	35 - 62 %	65 - 75 %
Abbildung: 7.5	$T_{60} \approx 0,7\ s$	SMD = 400 cm	38 - 58 %	68 - 81 %

7.4.2.3 Abhängigkeit von T_{60} bei konstantem SMD

Wie bereits bemerkt ist das Ermitteln von Systemverhalten als Abhängigkeit von der Nachhallzeit meist der einzige durchgeführte Test, um Systeme unter Hallbedingungen zu evaluieren. Allerdings findet man meist keine Angabe zum SMD.

In diesem Abschnitt wird ebenfalls das Verhalten der FCR in Abhängigkeit von der Nachhallzeit getestet. Als Besonderheit besteht hier die Unterscheidung zwischen Nahfeldverhalten und Fernfeldverhalten. Als Nahfeld wird, wie in den vorhergehenden Kapiteln, für die Tests der feste Abstand $r_{\text{Test}} = 1$ m festgelegt. Als Fernfeldabstand wird $r_{\text{Test}} = 3$ m festgelegt.

Die Ergebnisse für das Fernfeld (Abbildung 7.6 (unten)) spiegeln in etwa die Ergebnisse aus dem Experiment mit den künstlichen RIRs wider (Abbildung 7.2 (unten)).

Die Ergebnisse für FCR im Nahfeld für den Fall weibliche Sprache (Abbildung 7.6 (oben,rechts)) zeigen einen geringen Abfall bei steigender T_{60} bis etwa $T_{60} = 0,7$ s, gefolgt von einem nahezu konstanten Verlauf. Für den Fall männliche Sprache ist ein ähnliches Verhalten zu beobachten, allerdings sind die Ergebnisse, wie erwartet, generell schlechter.

Wenn man die Ergebnisse der Untersuchungen bei Fernfelddistanz und Nahfelddistanz bei der Nachhallzeit des SMART-Rooms ($T_{60} = 0.7$ s) mit den Ergebnissen bei den entsprechenden SMDs in Abbildung 7.5 (oben) vergleicht, können ähnliche Ergebnisse für den weiblichen Fall festgestellt werden. Für den Fall männliche Sprache lassen sich die Ergebnisse nicht abbilden. Die Unterschiede, die trotz gleicher Nachhallzeit auftreten (wie bereits in ASR-Experimenten Kapitel 6), sind mit den unterschiedlichen akustischen Eigenschaften von Räumen zu erklären.

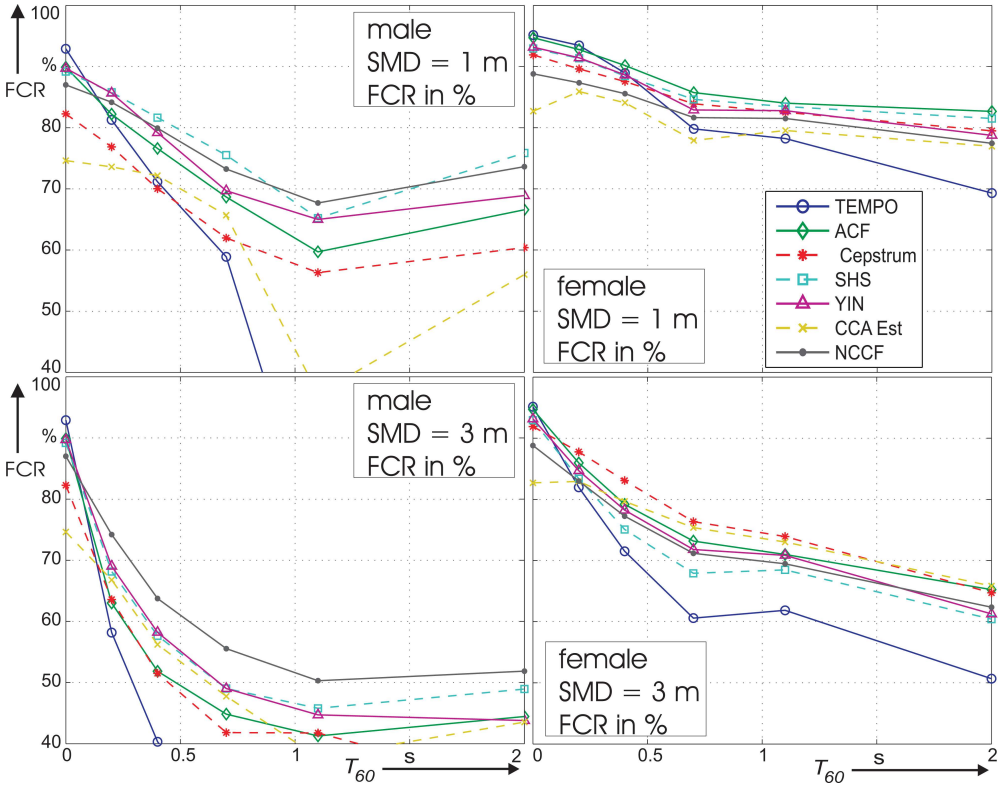

Abbildung 7.6 – Abhängigkeit der FCR von T_{60} bei SMDs im Nahfeld $r_{\mathrm{Test}} = 1$ m (oben) und im Fernfeld $r_{\mathrm{Test}} = 3$ m (unten). Links: männliche Sprecher, rechts: weibliche Sprecher.

7.5 Postprocessing-Verfahren

Der Basic-Extraktor generiert einen oder mehrere F_0-Kandidaten pro Frame. D. h., auch in einem stimmlosen Abschnitt entstehen zunächst F_0-Kandidaten. Der Postprocessor hat die Aufgabe, die Rohkandidaten des Basic-Extraktors auf Fehler zu untersuchen und diese möglichst zu beseitigen. Es entstehen zwei Arten von Fehlern mit den zugehörigen zwei Aufgaben: a) Beseitigen von Fehlern in der F_0-Kontur, b) Stimmlos-Klassifikation von F_0-Kandidaten in stimmlosen Abschnitten.

7.5.1 Überblick

Die im Folgenden benannten Ansätze sind in der Literatur bekannt.

Glättung der F_0-Kontur Ansätze, die die erste Aufgabe bearbeiten, beruhen auf der Tatsache, dass die F_0-Kontur sich nicht sprungartig ändern kann. Deshalb liegt es nahe, zunächst durch Glättung der F_0-Kontur sporadische Ausreißer zu unterdrücken. Die Glättung kann linear (Tiefpassfilter [Hes08]) und nichtlinear (Medianfilter [RSS75]) erfolgen. Es wurde gezeigt, dass Glättung die Ergebnisse fast aller getesteten Basic-Extraktor-Methoden erheblich gesteigert hat, z. B. [RCRM76, Spe84]. Glättungsverfahren sind nicht in der Lage Aufgabe b) zu lösen. Dazu wird ein zusätzlicher VUD benötigt (vgl. Abschnitt 7.6.3).

Korrektur durch dynamische Programmierung Eine weitere wichtige Postprocessing-Methode ist die dynamische Programmierung (DP) [Tal95, SD82]. Sie basiert darauf, dass der Basic-Extraktor eine Anzahl N_{cand} ($N_{\text{cand}} = 2 \ldots 5$) von F_0-Kandidaten $F_{0,\text{cand}}$ pro Frame generiert. D. h., es werden nicht nur das größte Maximum, sondern auch die nächstfolgenden Maxima in der Analysefunktion $A(p)$ gesucht (vgl. Gleichung (7.10)). Im Sinne der DP bilden die Kandidaten Zustände pro Zeiteinheit a (Zustandsindizes: i – gehört zum Zeitpunkt $(a-1)$, j – gehört zum aktuellen Zeitpunkt a, $F_{0,j}(a) = F_{0,i}(a-1)$, $1 \leq (i,j) \leq N_{\text{cand}}$). Nun wird der optimale Pfad durch die Zustände gesucht (graphische Erläuterung in [Hof98, QSS02]). Einige typische Implementierungen der DP bilden eine Kostenfunktion $C_{\text{ges}}(a)$, welche positive und negative Kosten C bis zum Zeitpunkt a aufakkumuliert. Der Pfad mit dem geringsten $C_{\text{ges}}(a)$ ist letztlich der Gewinner. Ein Beispiel für positive Kosten (Strafgewichte) ist dabei die Höhe der F_0-Differenz beim Übergang vom i-ten zum j-ten Zustand (engl.: Transition Costs)

$$C_{i,j}(a) = |F_{0,i}(a) - F_{0,j}(a)|. \tag{7.36}$$

Um dieses Beispiel zu verallgemeinern (unterschiedliche Implementierungen der DP), sollen die Übergangskosten durch eine Wichtungsfunktion w abhängig von den F_0-Kandidaten

$$C_{i,j}(a) = w(F_{0,i}(a), F_{0,j}(a)) \tag{7.37}$$

ausgedrückt werden. Als negative Kosten, also begünstigend wirkend, wird z. B. die Stärke der Maxima in der ACF oder im Cepstrum herangezogen. Sie sagt bekanntlich aus, wie signifikant F_0 gemessen werden kann. Die negativen Kosten werden auf verschiedene Weise in die Wegberechnung einbezogen (hier keine genaueren Angaben, vgl. Literatur, z. B. [LSN+07]).

Durch die Wegsuche hat die DP die Möglichkeit, auch dem zweit- oder drittbesten F_0-Kandidaten den Vorrang zu geben, auch wenn sich der erste Kandidat deutlicher in der $A(p)$ abhebt. Diese Vorgehensweise ist offensichtlich genauer als die Korrektur eines Ausreißers durch Glättung, weshalb die DP auch bessere Ergebnisse liefert [Tal95, SD82].

Ein weiterer Vorteil gegenüber Glättungsmethoden ist die VUD-Fähigkeit der DP. Dafür setzt man einfach einen nullten F_0-Kandidaten (Stimmloszustand) auf die Frequenz $F_{0,0} = 0$ Hz. Es entstehen die vier typischen Möglichkeiten der Zustandsübergänge:

Zustandsübergang	Abkürzung	Kostenfunktion
stimmhaft-stimmhaft	V-V	$w(F_{0,i}(a), F_{0,j}(a))$
stimmhaft-stimmlos	V-U	w_{VU}
stimmlos-stimmhaft	U-V	w_{UV}
stimmlos-stimmlos	U-U	w_{UU}

In stimmlosen Abschnitten erzeugt der Basic-Extraktor zufällig generierte Werte für F_0, sodass die F_0-Kontur zunächst einem Rauschen ähnelt. Für einen solchen Verlauf müssen die Gewichte w_{VU}, w_{UV} und w_{UU} (es sind Konstanten, wobei auch hier Funktionen denkbar sind) so eingestellt werden, dass die DP den Weg über den Zustand $F_{0,0}$ am günstigsten bewertet (VUD-Entscheidung). Zur Einstellung der Gewichte findet man in der Literatur keine Angaben, es handelt sich um ein heuristisches Problem.

DP wird u. a. in der Praat-Software [Boe93] eingesetzt. [QSS02, LSN+07] u. a. zeigen, dass die DP besonders robustheitssteigernd unter Geräuschbedingungen wirkt.

Im Rahmen dieser Arbeit wurden zwei Verfahren implementiert:

- **DP** nach [LSN+07] in Form des Viterbi-Algorithmus. Einzelheiten sind [LSN+07] zu entnehmen. Dies gilt, wie erwähnt, nicht für die Gewichte, zu deren Einstellung Experimente durchgeführt wurden.
- **Impulsunterdrückung** nach einer Eigenentwicklung. Eine genaue Beschreibung wird im folgenden Abschnitt dargestellt.

Beide Postprocessingverfahren wurden getestet. Die rechnerisch aufwendigere DP zeigte in den Experimenten gegenüber der Impulsunterdrückung keinen signifikanten funktionalen Vorteil. Dies entspricht in etwa auch den Ergebnissen in [QSS02], wo die DP nach ACF nur eine geringe Steigerung im Vergleich zur reinen ACF zeigt. Die Impulsunterdrückung kann als simpel aber effektiv eingeschätzt werden. Eine genaue Evaluation zur Güte der Verfahren wird nicht vorgenommen.

7.5.2 Aufbau eines Postprocessors mit Impulsunterdrückung

Da die Impulsunterdrückung als Postprocessing-Maßnahme im HFA-Ansatz integriert ist, soll sie im Folgenden genauer beschrieben werden. Sie kann mit der Medianfilterung verglichen werden. Als Impuls wird definiert, wenn die F_0-Differenz zu Vorgänger- und Nachfolgerwerten der F_0-Kontur größer als ein Schwellwert $\Delta_{F_0,\mathrm{Thresh}}$ ist. Es wird ein $\Delta_{F_0,\mathrm{Thresh}} = b_v \cdot F_0(a-1)$ mit $b_v = 0,3$ eingestellt. Motiviert ist dieser Wert dadurch, dass Praktiker die Geschwindigkeit einer Stimmanhebung etwa mit einer

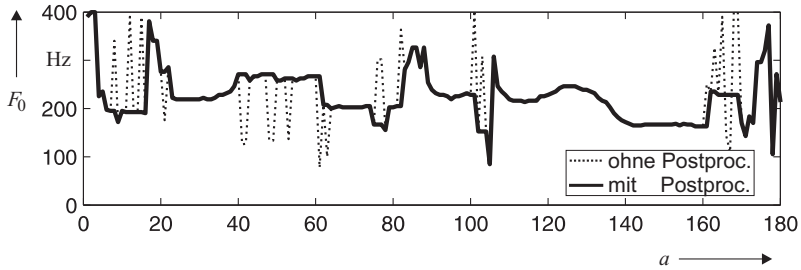

Abbildung 7.7 – Beispiel für eine F_0-Kontur generiert durch ACF im Vergleich mit der durch Postprocessing (Impulsunterdrückung) korrigierten F_0-Kontur. In Abschnitten in denen keine Impulse auftreten, befinden sich stimmhafte Bereiche, in denen F_0 deutlich messbar ist. In Abschnitten, wo besonders viele Impulse auftreten, befinden sich stimmlose oder schwache stimmhafte Bereiche.

Verdopplung in 35 ms (Maximalwert) bzw. mit 60 % in 35 ms (Durchschnittswert) beschreiben. Daraus ergibt sich bei einem Frameintervall $FI = 10$ ms ein $b_{v,\max} = 0,285$ bzw. $b_{v,\text{mean}} = 0,174$. Es sollen einzelne und doppelte Impulse entfernt werden (vgl. Abbildung 7.7). Drei oder mehr nacheinanderfolgende Werte, die aus einer F_0-Kontur heraus ragen, werden als Sprung behandelt. Ein Sprung wird nicht korrigiert, da er z. B. am Übergang von einem stimmlosen zu einem stimmhaften Abschnitt auftreten kann und dann eine korrekte Messung repräsentiert. In der Implementierung wird zunächst die Differenz $\Delta_{F_0,1} = |F_0(a) - F_0(a-1)|$ gebildet. Wenn diese größer als $\Delta_{F_0,\text{Thresh}}$ ist, wird anhand der Differenzen $\Delta_{F_0,2} = |F_0(a+1) - F_0(a-1)|$ und $\Delta_{F_0,3} = |F_0(a+2) - F_0(a-1)|$ entschieden, ob es sich um einen Sprung oder einen Impuls handelt. Wenn $\Delta_{F_0,3}$ kleiner als $\Delta_{F_0,\text{Thresh}}$ ist, handelt es sich bei den beiden Vorgängern je nach $\Delta_{F_0,2}$ um einen Einzel- oder Doppelimpuls, der korrigiert wird. Sind alle drei $\Delta_{F_0,i}$ größer als $\Delta_{F_0,\text{Thresh}}$, handelt es sich um einen Sprung, der nicht korrigiert wird. Die Korrekturwerte werden durch eine lineare Interpolation gebildet.

7.6 VUD-Verfahren

7.6.1 Überblick

In Abschnitt 7.2.1 wurden die Aufgaben des VUD bereits vorgestellt. Ansätze zu VUD-Verfahren lassen sich nach [Hes08] in drei Gruppen einteilen:

- **Schwellwertbasierte Verfahren** messen die Amplitude oder die Energie einer oder mehrerer charakteristischer Größen und vergleichen sie mit einem Schwellwert. Dabei ist es sinnvoll, eine Hysterese einzubeziehen, d. h., zwei Schwellwerte, je einen für das Ein- und einen für das Ausschalten, festzulegen. Charakteristi-

sche Größen für stimmhafte Signalabschnitte können z. B. die Impulshöhen in der ACF oder im Cepstrum sein. Auch die reine Energiekurve des Signals oder von Signalsubbändern kann in ihrer Form als charakteristische Größe betrachtet werden. Die Einstellung des Schwellwertes sollte möglichst adaptiv erfolgen, z. B. im Verhältnis zur Signalenergie oder SNR. Trotz adaptiver Einstellung gilt es auch hier, Heuristiken zur Einstellung von Konstanten anzuwenden.

- **Statistische Klassifikatoren** treffen die VUD-Entscheidung anhand eines trainierten Modelles. Zum Training und zur Erkennung werden Merkmale, gewonnen aus charakteristischen Größen, verwendet. Die Merkmale können ein- oder mehrdimensional sein, d. h., die parallele Beobachtung mehrerer charakteristischer Größen ist denkbar. Im einfachsten Fall benutzt man einen Bayes-Klassifikator, aufwendiger ist ein HMM, welches auch phonembasiert sein kann. Statistische VUD-Klassifikatoren wurden in dieser Arbeit nicht untersucht.
- **Integrierter Ansatz** bezeichnet Verfahren, wo VUD und F_0-Detektion miteinander verschmolzen sind. Dazu zählt insbesondere das im vorigen Abschnitt beschriebene DP-Verfahren.

Unabhängig von der VUD-Methode kann davon ausgegangen werden, dass die VUD-Entscheidung unsicher wird, sobald Sprache gestört ist. Im Falle von Geräuschstörung würde ohne Zusatzmaßnahmen die Anzahl von Voiced Errors N_{VE} steigen, da die F_0-Komponente im Signal, ausgedrückt durch eine charakteristische Größe, am Übergang von V-U oder U-V schwächer wird und je nach SNR im Rauschen untergeht. Im Falle einer Hallstörung ist die Aufgabe weitaus schwieriger. Der Beginn eines stimmhaften Abschnittes ist vergleichsweise gut zu erkennen. Das Ende stellt jedoch ein großes Problem dar. Es wird nicht nur die Energie des vorigen stimmhaften Lautes (vgl. Abbildung 7.8), sondern auch seine harmonische Struktur in den Bereich des stimmlosen verschleppt (vgl. Abbildung 4.4). Deshalb erscheint im Bereich des stimmlosen Lautes das Signal immer noch mit Merkmalen eines stimmhaften Lautes, wodurch die Aufgabe für den VUD sehr schwierig wird. Im Rahmen dieser Arbeit wurden zwei Verfahren implementiert:

- **Integrierter Ansatz** in Form der DP nach [LSN+07]. In diesem Verfahren sind Postprocessor und VUD integriert (Beschreibung vgl. Abschnitt 7.5).
- **Schwellwertbasierter Ansatz** nach einer Eigenentwicklung. Eine genaue Beschreibung erfolgt im nächsten Abschnitt.

Aufgrund von Einzeltests und der geringeren Komplexität fiel die Entscheidung auf den schwellwertbasierten Ansatz. Der integrierte Ansatz wurde nicht umfangreich untersucht.

7.6.2 Aufbau eines schwellwertbasierten VUD

In Abschnitt 6.3.2 wird beschrieben, wie aufgrund der F_0-Kontur für jeden Frame a das Spektrum der Harmonischen $X_h(a,n)$ berechnet wird. Für dieses VUD-Verfahren wird aus $X_h(a,n)$ die mittlere Energie der Harmonischen

$$E_h(a) = \frac{1}{i_{\max}(a)} \sum_{i=1}^{i_{\max}(a)} X_h^2(a, n_{h,i}(a)) \tag{7.38}$$

bestimmt. Es werden Harmonische bis zu einem framespezifischen maximalen Vielfachen $i_{\max}(a)$ von F_0 in die Berechnung einbezogen. Zur Berechnung von $i_{\max}(a)$ siehe Gleichung (6.6). In stimmlosen Abschnitten a_u kann $E_h(a_u) \approx E(a_u)$ angenommen werden ($E(a)$ – mittlere Energie des Gesamtspektrums). $E_h(a_u)$ basiert hier auf einem zufälligen, fiktiven F_0, das vom Basic-Extraktor bereitgestellt wird. Die Idee besteht darin, dass sich die Harmonischen in stimmhaften Frames a_v deutlich aus dem Spektrum herausheben. Das bedeutet $E_h(a_v) > E(a_v)$. Die Kontur von $E_h(a)$ stellt demnach stimmhafte Abschnitte deutlicher heraus. Letztlich wird $E_h(a)$ mit einem Schwellwert

$$E_{th} = \frac{b_{th}}{A} \sum_{a=0}^{A-1} E_h(a) \tag{7.39}$$

verglichen, der sich aus dem Mittelwert über alle E_h einer Äußerung ergibt (vgl. Abbildung 7.8, A – Anzahl Frames der aktuellen Äußerung). Der Parameter b_{th} ist dabei ein Proportionalitätsfaktor, der die Sensibilität des VUDs einstellt. Er wird zunächst fest eingestellt, wobei dieser Wert im Rahmen dieser Arbeit umfangreich optimiert wurde (Abschnitt 7.6.3 für die VUD-Leistung, Abschnitt 6.4.3 für die ASR-Leistung). Diese adaptive Methode der Berechnung des Schwellwertes E_{th} ist zunächst simpel, aber, wie in Experimenten nachgewiesen wird, wirkungsvoll. In einer realen Anwendung der Kommandoworterkennung bietet diese VUD-Methode eine ausreichende Qualität (ein robust arbeitender VAD wird vorausgesetzt; das Thema VAD wird in dieser Arbeit nicht behandelt). Grund dafür ist, dass am Ende des Kommandos die Schwellwertberechnung, gefolgt von den weiteren ASR-Algorithmen, stattfinden kann, die letztlich zur Klassifikation nach Beendigung der Äußerung führen. Die Berechnungsmethode von E_{th} kann als ungünstig betrachtet werden, wenn es sich um eine kontinuierliche Erkennung von Sätzen, Texten oder Spontansprache handelt (z. B. Diktiersystem). Für diese Anwendungsbereiche empfiehlt der Autor, die Berechnung von E_{th} nicht über die gesamte Äußerung durchzuführen, sondern in kurze Signalabschnitte, z. B. $A = 50$ ($\hat{=}$ 500 ms), aufzuteilen. Weiterhin wird ein Überlappen der Signalabschnitte sowie eine Glättung der entstehenden E_{th}-Kurve vorgeschlagen. Eine solche kontinuierliche Berechnung von E_{th} würde zur Echtzeitfähigkeit des VUD für längere Äußerungen führen; dies wurde im Rahmen dieser Arbeit nicht getestet.

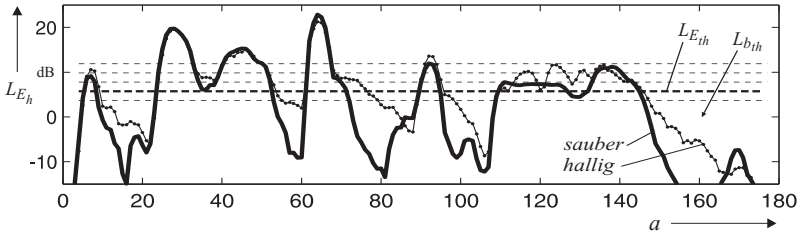

Abbildung 7.8 – Mittlere Energie der Harmonischen in Pegeldarstellung $L_{E_h} = 10 \lg E_h(a)$ dB für eine Äußerung. Fette Linie: sauberes Signal; dünne Linie: verhalltes Signal mit $T_{60} = 700$ ms, SMD = 1 m. Die fette gestrichelte Linie beschreibt den Schwellwert in Pegeldarstellung $L_{E_{th}} = 10 \lg E_{th}$ dB für $b_{th} = 0{,}25$. Die beiden Schwellwerte für das saubere und das verhallte Signal unterscheiden sich in diesem Beispiel um 0,5 dB; deshalb wird aus Darstellungsgründen für beide Werte nur eine Linie abgebildet. Die Schar von dünnen gestrichelten Linien beschreibt den Einfluss von b_{th}, durch den sich der Schwellwert $L_{E_{th}}$ additiv nach oben oder unten verschieben lässt. In diesem Beispiel beträgt der Unterschied zwischen den einzelnen Verläufen von $L_{E_{th}}$ jeweils 2 dB; gleiches gilt für die zugehörigen Einstellungen von $L_{b_{th}} = 10 \lg b_{th}$ dB.

7.6.3 Evaluation des schwellwertbasierten VUD unter Hallbedingungen

Im Gegensatz zur vergleichenden Evaluation der Basic-Extraktoren untersuchen die folgenden Experimente nur eine VUD-Methode, den schwellwertbasierten VUD, wie er im vorangegangenen Abschnitt vorgestellt wurde. Die untersuchten Hallabhängigkeiten sind die gleichen wie sie in der Evaluation der Basic-Extraktoren verwendet wurden (die vier Hallabhängigkeiten (a) ... (d) aus Abschnitt 7.3.3). Die Experimente wurden mit der Datenbasis **(iii)** (ECESS-Datenbasis, vgl. Abschnitt 7.3.1) durchgeführt.

7.6.3.1 UER und VER in Abhängigkeit vom SMD

Zunächst wurde ein einführendes Experiment durchgeführt, welches das Verhalten der Fehlerraten UER und VER (vgl. Gleichungen (7.5) und (7.6)) bei Variation von b_{th} untersucht. Abbildung 7.8 beschreibt graphisch, welchen Einfluss b_{th} auf das VUD-Verhalten hat. Ist b_{th} zu groß, werden zu wenige stimmhafte Frames korrekt klassifiziert, VER steigt. Im Umkehrschluss werden mehr und mehr stimmlose Frames korrekt klassifiziert, UER sinkt, letztlich bis auf 0. Wird hingegen b_{th} zu klein gewählt, entsteht ein umgekehrtes Verhalten: mehr und mehr Frames werden als stimmhaft klassifiziert, VER sinkt, und zu wenige stimmlose Frames werden korrekt klassifiziert, UER steigt. Auf dieses Verhalten kann man bereits durch Überlegung schließen. Die Experimente, die die entsprechenden Raten unter Hallbedingungen messen, weisen zunächst diese Überlegung nach, vgl. Abbildung 7.9. Es wurde nur die Verhallung in Abhängigkeit

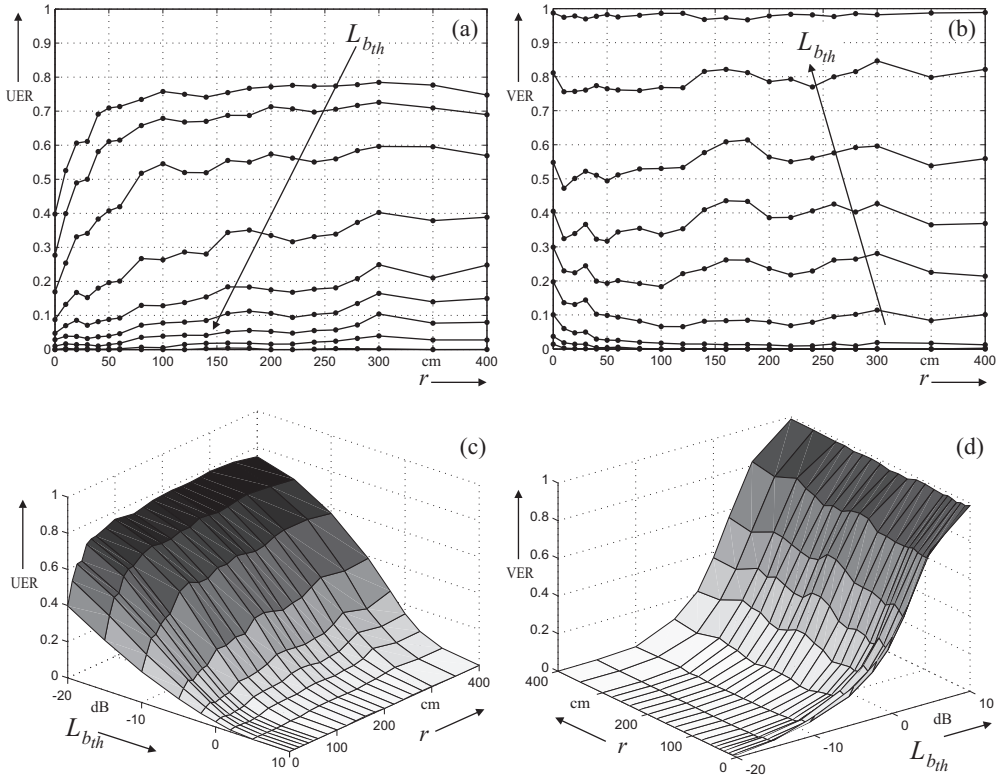

Abbildung 7.9 – VER und UER in Abhängigkeit vom SMD (SMART-Room) bei verschiedenen Einstellungen von $L_{b_{th}} = (-20; -15; -10; -5; -2; 0; 2; 5; 10)$ dB. (a) UER 2D-Darstellung, (b) VER 2D-Darstellung, (c) UER 3D-Darstellung und (d) VER 3D-Darstellung.

vom SMD (vgl. Abschnitt 7.3.3) verwendet. Ein interessantes Ergebnis ist, dass VER nahezu unabhängig vom Grad der Verhallung ist, wohingegen bei UER eine starke Abhängigkeit festzustellen ist (innerhalb des Hallradius).

7.6.3.2 ER – Gesamtverhalten des VUDs

Die Fehlerraten UER und VER sagen einzeln gemessen noch nicht genug über das VUD-Verhalten aus. Um beide Fehlerraten zu kombinieren, wird die Gesamtfehlerrate ER nach Gleichung (7.8) gebildet. Es ist zu erwarten, dass ein oder mehrere Werte für b_{th} existieren, bei denen der VUD optimal arbeitet, d. h., ER minimal wird. Um dies zu untersuchen, wurde ER für alle vier evaluierten Hallabhängigkeiten ermittelt und

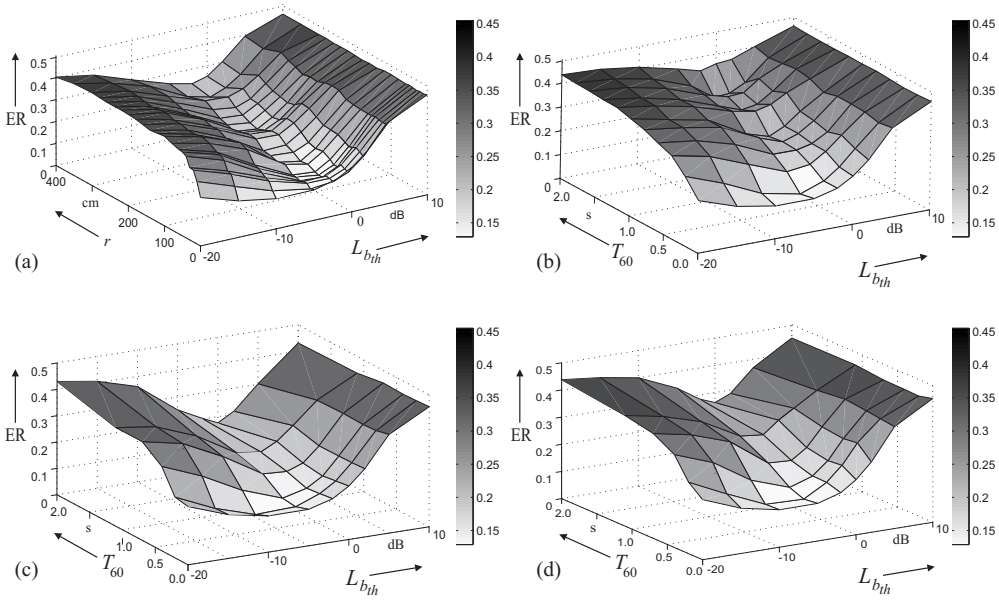

Abbildung 7.10 – VUD-Gesamtfehlerrate ER abhängig vom logarithmierten Threshold-Parameter $L_{b_{th}}$ bei den vier verschiedenen Hallabhängigkeiten: (a) SMD (SMART-Room), (b) künstliche Verhallung, (c) T_{60} bei SMD = 1 m und (d) T_{60} bei SMD = 3 m.

in Abbildung 7.10 dargestellt.

Zunächst ist zu erkennen, dass auch bei sauberen Daten im besten Fall nur ca. 15 % der Frames falsch klassifiziert werden. Dieses Ergebnis beschreibt die maximale Leistungsfähigkeit des untersuchten VUDs. Zum Vergleich liefern die Ergebnisse der drei untersuchten PDA-Algorithmen mit integrierter VUD-Funktionalität in der ECESS-Evaluation für saubere Daten ebenfalls Ergebnisse von ER = (10 ... 20) % [Hoe07]. Bei gestörten Umgebungen steigen die Fehlerraten erheblich an, [Hoe07] berichtet von ER = (25 ... 60) % je nach Verfahren und Geräuschtyp. Für verhallte Daten liefert der hier vorgestellte VUD bei optimaler Einstellung von b_{th} Ergebnisse von ER = (25 ... 35) %. Die Graphiken beschreiben weiterhin, dass das optimale b_{th} bei stärkerer Verhallung ansteigt. So liegt es für saubere Daten bei $L_{b_{th}} = (-10 \ldots -5)$ dB. Bereits bei geringer Verhallung liegt das Optimum etwa bei $L_{b_{th}} = -2$ dB und tendiert bei stärkerer Verhallung zu $L_{b_{th}} = 0$ dB. Dieses Verhalten erklärt sich graphisch in Abbildung 7.8. Ein Vergleich der Ergebnisse mit anderen Ansätzen ist nicht möglich, da derzeit keine weitere Untersuchung von VUD-Verfahren unter Hallbedingungen veröffentlicht wurde.

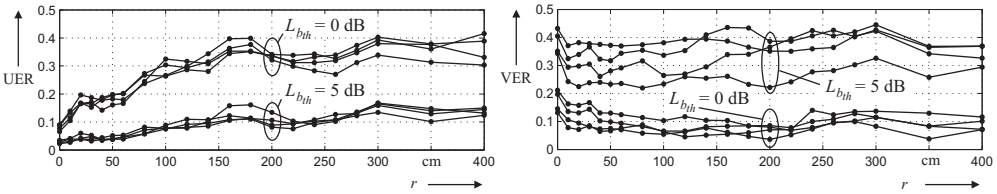

Abbildung 7.11 – Darstellung der Unabhängigkeit von UER und VER von der Sprache und des Geschlechts der Sprecher. Die vier Graphen in einem Graphenbündel enthalten die Fehlerraten für männlich/weiblich jeweils in englisch/deutsch. Aus Darstellungsgründen wurden nur die zwei Beispiele für $L_{b_{th}}$ von 0 dB und 5 dB ausgewählt.

7.6.3.3 Unabhängigkeit von Sprache und Geschlecht

Eine im Vorfeld durchgeführte Untersuchung ergab das in Abbildung 7.11 dargestellte Ergebniss, dass das VUD-Verhalten des hier untersuchten schwellenbasierten VUDs weder vom Geschlecht, noch von der Sprache abhängig ist. Der Grund dafür ist, dass dieses VUD-Verfahren relativ unabhängig von der Güte der F_0-Detektion ist. Bei einem VUD, der z. B. auf der Deutlichkeit der Periodizität in der ACF ($\psi_{xx}(F_0)/\psi_{xx}(0)$, vgl. ACF) basiert, wäre anzunehmen, dass dem nicht so ist. Aufgrund dieser Erkenntnis wurde in den Abbildungen 7.9 sowie 7.10 aus den vier Kombinationen der Sprechercharakteristiken männlich/weiblich, englisch/deutsch deshalb nur die Variante weiblich-englisch dargestellt.

7.7 Zusammenfassung der Erkenntnisse

In diesem Kapitel wurde das Verhalten von PDA-, Postprocessing- und VUD-Methoden unter Hallbedingungen untersucht. Folgende Erkenntnisse konnten gewonnen werden:

- Die Performanz aller üblichen PDA-Verfahren sinkt drastisch bei Vorhandensein von Hall. Es existiert derzeit kein Verfahren, welches in der Lage ist, robust unter Hallbedingungen zu arbeiten. Am besten schnitten die Verfahren ACF, NCCF, Cepstrum, SHS und YIN ab. Keines dieser Verfahren arbeitet signifikant besser oder schlechter als die anderen.
- PDA-Verfahren arbeiten unter Hallbedingungen besser für weibliche als für männliche Sprache. Eine Erklärung dafür wird in Abschnitt 7.4.2.1 gegeben.
- Da die Harmonischen aus F_0 berechnet werden müssen, ist es für die Anwendung in HFA sinnvoll, die Performanz von PDA-Algorithmen in FCR anzugeben (im Gegensatz zum üblichen GER). Bei Verhallungen in typischen Wohn- und Büroumgebungen erreichen die benannten Algorithmen im Fernfeld etwa FCR = (50 … 70) % für männliche Sprache und FCR = (75 … 85) % für

weibliche Sprache. Im Nahfeld steigern sich die Resultate.

- Für den Einsatz innerhalb der HFA-Methode wird ein Postprocessing-Verfahren der Impulsunterdrückung vorgestellt. In Einzeltests können keine Vor- oder Nachteile gegenüber der dynamischen Programmierung festgestellt werden. Deshalb wird dem Verfahren aufgrund der geringeren Komplexität der Vorzug gegenüber der DP gegeben.
- Für den Einsatz innerhalb der HFA-Methode wird ein schwellwertbasiertes VUD-Verfahren vorgestellt. Es arbeitet auf sauberen Daten ähnlich gut wie andere Verfahren, deren Evaluation mit der gleichen Datenbasis in der Literatur beschrieben ist.
- Das vorgestellte VUD-Verfahren wird umfangreich unter allen experimentellen Hallbedingungen getestet. Der optimale Wertebereich für den VUD-Schwellwertparameter liegt bei $L_{b_{th}} = (-5 \ldots 0)$ dB. Bei Verhallungen in typischen Wohn- und Büroumgebungen erreicht der vorgestellte VUD Fehlerraten zwischen etwa ER $= (15 \ldots 25)$ %.
- Die Performanz des vorgestellten VUD-Verfahrens ist nicht vom Geschlecht der Sprecher oder der Sprache abhängig.

8 Zusammenfassung und Ausblick

Die vorliegende Arbeit hat den Anwendungshintergrund der Sprachsteuerung für elektronische Geräte im Haushaltsbereich (Wohn- und Büroumfeld). Dabei entstehen sowohl rein praktische als auch wissenschaftlich relevante Probleme. Herkömmlich werden Spracherkenner mit Merkmalmustern ungestörter Trainingsdaten trainiert. In realen akustischen Umgebungsbedingungen unterscheiden sich die Merkmalmuster der Sprachsignale durch das Vorhandensein zusätzlicher Störanteile. Die Abweichung von Trainings- und Testmustern führt zu einer Verschlechterung der Erkennungsleistung. Es existieren die beiden Störungsarten der additiven (Rauschen, konkurrierende Sprecher, Türenklappen etc.) und der convolutiven Störungen (Kanalverzerrungen und Raumhall). Während Lösungsansätze für additive Störungen bereits weitreichend untersucht sind, widmet sich diese Arbeit dem noch nahezu ungelösten Problem der raumakustischen Störungen, dem Raumhall. In dieser Arbeit wird angenommen, dass additive Störungen nicht vorhanden sind, sodass der Raumhall die einzige Störgröße darstellt.

Nach einer Einleitung und der Vorstellung einer Experimentierumgebung beginnt die Arbeit in Kapitel 3 mit der Beschreibung der raumakustischen Umgebungsbedingungen. Die Darstellung erfolgt zunächst anhand der physikalischen Prinzipien der Raumakustik, wobei typische Eigenschaften von Schallfeldern in Räumen beleuchtet werden. In einem Übergang von der physikalischen zu einer systemtheoretischen Betrachtung der Sprecher-Raum-Mikrofon-Strecke (SRM-System) werden die Raumimpulsantwort (RIR) als wichtigste Beschreibungsform des SRM-Systems mathematisch modelliert und wesentliche Eigenschaften abgeleitet. Besonders hervorzuheben ist dabei, dass der Hall, der die Störung beschreibt und im Raum (idealisiert betrachtet) gleichverteilt ist, und der Direktschall, der das Nutzsignal darstellt und umgekehrt proportional vom Sprecher-Mikrofon-Abstand (SMD) abhängig ist, sich additiv überlagern. Die Störung ist damit nicht nur von der Halligkeit des Raumes, die durch die Nachhallzeit T_{60} beschrieben wird, abhängig, sondern auch vom SMD. Um die Grenzen der Hallstörungen im Wohn- und Büroumfeld zu untersuchen, wird eine statistische Untersuchung durchgeführt, die als Ergebnis eine Nachhallzeit zwischen $T_{60} = (0,3 \ \ldots \ 0,8)$ s und SMDs zwischen $r = (0,5 \ \ldots \ 4)$ m in derartigen Umgebungen feststellt. Zusätzlich wird in der Tendenz nachgewiesen, dass T_{60} mit steigender Frequenz geringer wird, auch wenn das für einzelne Räume eventuell nicht zutreffen sollte.

Nach den Ausführungen zu akustischen Umgebungsbedingungen verfolgt Kapitel 4 das Ziel, die Störung der Sprache durch raumakustische Effekte zu bestimmen so-

wie deren Auswirkungen auf die Spracherkennung zu untersuchen. Das Kapitel beginnt mit der einführenden Beschreibung wichtiger Aspekte ungestörter menschlicher Sprachsignale. Es folgt eine Beschreibung von verhallten Sprachsignalen, die sich in die Beschreibung im Signalbereich und die Beschreibung im Modulationsbereich der zeitlichen Einhüllenden (TPEs) von Sprachsignalen unterscheidet. Es wird die Modulationsübertragungsfunktion (MTF) eingeführt, die beschreibt, wie sich Raumhall auf die TPEs auswirkt. Nach diesen theoretischen Betrachtungen wird versucht, den Grad der Störung durch Raumhall quantitativ festzustellen. Dabei erfolgt ein kurzer Überblick über traditionelle Maße, die subjektiv oder objektiv die Störung durch den Hall beschreiben sollen. Die wichtigsten Störmaße (z. B. Deutlichkeitsmaß C_{50}) werden in einem Experiment mit den später gemessenen Erkennungsraten unter Hallbedingungen verglichen. Es stellt sich heraus, dass die einschlägigen Maße ungeeignet sind, um die Störung der Spracherkennung durch Hall zu beschreiben. Aufgrund dessen wird in dieser Arbeit ein Vorschlag für ein Störmaß erarbeitet, dessen Verlauf mit dem Verlauf der Erkennungsrate korrespondiert. Diese Untersuchungen erfolgen im Zusammenhang mit den ersten Spracherkennungsexperimenten, die das Verhalten des hier verwendeten Kommandoworterkenners in Abhängigkeit der beiden störungsbeschreibenden Parameter, T_{60} und SMD, überprüfen sollen. Es kann festgestellt werden, dass eine starke Abhängigkeit der Erkennungsrate von der Nachhallzeit, aber auch vom SMD vorliegt. Bereits bei geringen Verhallungen wird ein rapider Einbruch der Erkennungsrate festgestellt, auch für relativ geringe Hallstörungen, wie sie im Wohn- und Büroumfeld vorkommen. Um die Abhängigkeit des Erkenners von der Hallstörung weiter zu untersuchen, werden Experimente mit modifizierten RIRs durchgeführt, bei denen die Hallphase so beschnitten ist, dass die Wirkung einzelner Hallabschnitte ermittelt werden kann. Im Ergebnis kann insbesondere festgestellt werden, dass hochfrequenter Hall (> 2500 Hz) die Spracherkennung kaum stört, wohingegen tieffrequenter Hall (< 2500 Hz) besonders störend wirkt.

Die bis hier durchgeführten theoretischen und experimentellen Betrachtungen beschreiben die Problemstellung der Hallstörung umfangreich und gewinnen bereits neue wissenschaftliche Erkenntnisse. Das anschließende Kapitel 5 stellt die existierenden Lösungsansätze systematisch vor, die bislang für die Problemstellung der Hallstörungen entwickelt wurden. Da das Thema Hallstörungen für Spracherkenner noch recht jung ist, existiert noch keine einheitliche Herangehensweise. Der Autor lehnt sich, wie auch wenige ähnliche Beiträge, in Bezug auf die Einteilung der Maßnahmen gegen Raumhall an die Einteilung von Maßnahmen gegen additive Störungen an und stellt die vier Ebenen akustische Ebene, Signalebene, Merkmalebene sowie Modellebene als Ansatzpunkte für robustheitssteigernde Maßnahmen fest. Nach Vorstellung aller Ansätze wird deutlich, dass derzeit kein Ansatz existiert, der das Problem des Raumhalls für die Spracherkennung zufriedenstellend und unter Beachtung der praktischen Randbedingungen, die in der Einleitung (Abschnitt 1.2) vorgestellt werden, lösen kann.

Aufgrund dessen wird im Rahmen der Arbeit versucht, einen neuen Ansatz zu entwickeln, der von vornherein in Hinblick auf praktische Einsatzbedingungen ausgelegt ist. Dabei entsteht in Kapitel 6 die Methode Harmonicity-based Feature Analysis (HFA), die auf den drei Ideen basiert, dass harmonische Komponenten im Spektrum als ungestört, tiefe Frequenzen in stimmlosen Abschnitten besonders störend sowie hochfrequente Hallkomponenten als harmlos angenommen werden können. Die Methode versucht dabei sichere (als ungestört angenommene) Merkmale im Sprachsignal zu verstärken sowie unsichere (als gestört angenommene) Merkmale zu unterdrücken. Die HFA-Methode wird umfangreich unter systematischer Variation verschiedener Hallbedingungen getestet. Sie wird mit weiteren aus der Literatur bekannten Methoden verglichen und kombiniert. Diese sind die herkömmliche Merkmalanalyse (CFA), verhalltes Training, Temporal Power Envelope Feature Analysis (TPEFA), IMTF-basierte Enthallung sowie der Delay-and-Sum Beamformer (DSB). Es kann festgestellt werden, dass das verhallte Training erwartungsgemäß die stärkste robustheitssteigernde Wirkung besitzt. Dennoch kann für CFA auch bei verhalltem Training eine Abhängigkeit der Erkennungsrate von der Nachhallzeit festgestellt werden. HFA kann die Erkennungsrate nicht nur steigern, sondern ist auch in der Lage, diese stabil für mehrere Umgebungsbedingungen aufrecht zu erhalten. Zur detaillierteren Beschreibung der umfangreichen Experimente und deren Erkenntnisse wird auf das Kapitel 6 verwiesen. Zusammenfassend wird festgestellt, dass mit der Methode HFA ein technischer Fortschritt erzielt worden ist, der sich sowohl durch das Erreichen von praktisch relevanten Erkennungsleistungen als auch durch die praktische Anwendbarkeit auszeichnet und sich damit von anderen Ansätzen unterscheidet.

Die Methode HFA greift auf zwei Basistechnologien der digitalen Sprachsignalverarbeitung zurück, die F_0-Detektion und die Stimmhaft-Stimmlos-Entscheidung (VUD). Beide Technologien sind bislang (bis auf eine Ausnahme) noch nicht unter Hallbedingungen untersucht worden. Kapitel 7 hat das Ziel, wichtige Methoden der F_0-Detektion und der VUD unter Hallbedingungen zu testen und aufgrund der Ergebnisse Vorschläge für den Einsatz in HFA zu formulieren. Nach der Vorstellung einer Evaluationsumgebung sowie einem Überblick über existierende Verfahren für die beiden Technologien zeigen Experimente jeweils erwartungsgemäß, dass die entsprechenden Ansätze wesentlich schlechter als im ungestörten Fall funktionieren. Die Untersuchung beider Technologien unter raumakustischen Umgebungsbedingungen stellt eine Neuerung dar. Aufgrund der Experimente kann je ein (ausreichend) erfolgreich arbeitendes Verfahren für die HFA-Methode festgestellt werden. Dabei handelt es sich im Falle der F_0-Detektion um den relativ simplen Ansatz der Autokorrelationsfunktion, der ähnlich gut wie einige weitere Ansätze abschneidet, aber aufgrund seiner günstigen Implementierungseigenschaften und der geringen Rechenleistung ausgewählt wird. Für den VUD wird ein im Rahmen dieser Arbeit neu vorgestellter schwellenbasierter Ansatz ausgewählt und dessen Leistungsfähigkeit nachgeweisen.

Die vorliegende Arbeit leistet wesentliche Beiträge zur Weiterentwicklung des For-

schungsgebietes der raumakustischen Störungen in der Spracherkennung sowie in der Sprachsignalverarbeitung allgemein. Es konnten neue Ansätze gefunden und deren Leistungsfähigkeiten unter praktischen Einsatzbedingungen nachgewiesen werden. Wie die Arbeit an vielen Stellen gezeigt hat, bietet das Thema der raumakustischen Störungen in der Sprachsignalverarbeitung eine Fülle von weiteren Ansatzpunkten, die in dieser Arbeit nicht weiter untersucht wurden, jedoch großes Potential für neue Erkenntnisse besitzen. Dies ist insbesondere dadurch gekennzeichnet, dass die Bearbeitung des Themas bislang noch am Anfang steht. Die weitere Untersuchung ist Gegenstand zukünftiger Forschung.

Literaturverzeichnis

[Abd73] Abdel Alim, O.: Abhängigkeit der Zeit- und Registerdurchsichtigkeit von raumakustischen Parametern bei Musikdarbietungen. Dissertation, Technische Universität Dresden, 1973.

[AH96] Avendano C. und Hermansky, H.: Study on the dereverberation of speech based on temporal envelope filtering. Proc. ICSLP 1996, Philadelphia, USA, 1996, S. 889 – 892.

[Ahn75] Ahnert, W.: Einsatz elektroakustischer Hilfsmittel zur Räumlichkeitssteigerung, Schallverstärkung und Vermeidung der akustischen Rückkopplung. Dissertation, Technische Universität Dresden, 1975.

[AIK+00] Atake, Y.; Irino, T.; Kawahara, H.; Lu, J.; Nakamura, S. und Shikano, K.: Robust fundamental frequency estimation using instantaneous frequencies of harmonic components. Proc. ICSLP 2000, Beijing, China, 2000, S. 907 – 910.

[ANR74] Ahmed, N.; Natarajan, T. und Rao, K. R.: Discrete Cosine Transform. IEEE Transactions on Computers, Vol. C-23(1974), S. 90 – 93.

[Ara96] Arai, T.; Hermansky, H.; Pavel, M. und Avendano, C.: Intelligibility of speech with high-pass filtered time trajectories of spectral envelopes, Proc. ICSLP 1996, Philadelphia, USA, 1996, S. 2490 – 2493.

[Ara97] Arai, T. und Greenberg, S.: The temporal properties of spoken Japanese are similar to those of English. Proc. EUROSPEECH 1997, Rhodes, Griechenland, 1997, S. 1011 – 1014.

[AS35] Aigner, F. und Strutt, M. J. O.: On a physiological effect of several sources of sound on the ear and its consequences in architectural acoustics. Journal of the Acoustical Society of America, Vol. 6(1935), No. 3, S. 155 – 159.

[ASNH07] Abad, A.; Segura, C.; Nadeu, C.; Hernando, J.: Audio-based approaches to head orientation estimation in a smart-room. Proc. INTERSPEECH 2007, Antwerp, Belgien, 2007, S. 590 – 593.

[Ata74] Atal, B, S.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustical Society of America, Vol. 55(1974), No. 6, S. 1304 – 1312.

[Atl07] Atlas, L.: New Theory Insights for Modulation Analysis and Filtering of Speech. Tutorial at the INTERSPEECH 2007, Antwerp, Belgien, 2007.

[AY79] Ananthapadmanabha, T. V. und Yegnanarayana, B.: Epoch extraction from linear prediction residual for identification of the closed glottis interval. IEEE Transactions on Audio, Speech and Signal Processing, Vol. 27(1979), No. 4, S. 309 – 319.

[BAK04] Buchner, H.; Aichner, R. und Kellermann, W.: TRINICON: A Versatile Framework for Multichannel Blind Signal Processing. Proc. ICASSP 2004, Montreal, Canada, 2004, S. 889 – 892.

[BAK05] Buchner, H.; Aichner, R. und Kellermann, W.: Relation between blind system identification and convolutive blind source separation. Joint Workshop for Hands-Free Speech Communication and Microphone Arrays (HSCMA), Piscataway, USA, 2005, S. 3 – 4.

[BBE72] Bass, H. E.; Bauer, H.-J.; Evans, L. B.: Atmospheric Absorption of Sound: Analytical Expression. Journal of the Acoustic Sosiety of America, Vol. 52(1972), No. 3B, S. 821 – 825.

[BBK91] Bees, D.; Blostein, M. und Kabal, P.: Reverberant speech enhancement using cepstral processing. Proc. ICASSP 1991, Toronto, Canada, 1991, S. 977 – 980.

[BG00] Brandstein, M. S. und Griebel, S. M.: Nonlinear, model-based microphone array speech enhancement. In: Gay, S. L. und Benesty, J. (Eds.): Acoustic Signal Processing For Telecommunication. Kluwer Academic Publishers, 2000, S. 261 – 279.

[BHT+04] Bonafonte, A.; Hoege, H.; Tropf, H. S.; Moreno, A.; Heuvel, H.; Suendermann, D.; Zeigenhain, U.; Perez, J. und Kiss, I.: TC-STAR – TTS baselines and specifications. Deliverable no.: D8 2004. http://www.tc-star.org/.

[BK08] Brüel & Kjær GmbH: DIRAC – Software Dokumentation. http://www.bruelkjaer.de/, 2008.

[Boe93] Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proceedings, Vol. 17, University of Amsterdam, Institute of Phonetic Sciences, Amsterdam, Niederlande, 1993, S. 97 – 110.

[Bor53] Boré, G.: Kurzton-Meßverfahren zur punktweisen Ermittlung der Sprachverständlichkeit in lautsprecherbeschallten Räumen. Dissertation, Technische Hochschule Aachen, 1953.

[Bol79] Boll, S. F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 27(1979), No. 2, S. 113 – 120.

[BP99] Boré, G. und Peus, S.: Microphones – Methods of Operation and Type Examples. 4. Edition, Georg Neumann, Berlin, 1999.

[Bri98] Brixen, E. B.: Near field registration of the human voice: Spectral changes due to positions. AES 104th Convention, 1998.

[BS65] Beranek, L. L. und Schultz, T.: Some recent experiences in the design and testing of concert halls with unspended panel arrays. Acustica 16(1965), S. 307.

[BSH08a] Benesty, J; Sondhi, M. M.; Huang, Y. (Eds.): Springer Handbook of Speech Processing. XXXVI, Springer-Verlag New York, 2008, ISBN 978-3-540-49125-5.

[BSH08b] Benesty, J; Sondhi, M. M.; Huang, Y.: Introduction to Speech Processing. In: Benesty, J; Sondhi, M. M.; Huang, Y. (Eds.): Springer Handbook of Speech Processing. XXXVI, Springer-Verlag New York, 2008, ISBN 978-3-540-49125-5, S. 1 – 4.

[BSK01] Bitzer, J.; Simmer, K. U. und Kammeyer, K. D.: Multi-microphone noise reduction techniques as front-end devices for speech recognition. Speech Communication, Vol. 34(2001), S. 3 – 12.

[BSZ+95] Bass, H. E.; Sutherland, L. C.; Zuckerwar, A. J.; Blackstock, D. T. und Hester, D. M.: Atmospheric absorption of sound: Further developments. Journal of the Acoustic Sosiety of America, Vol. 97(1995), No. 1, S. 680 – 683.

[BW01] Brandstein, M. und Ward, D.: Microphone Arrays: Signal Processing Techniques and Applications. Springer-Verlag, Berlin, 2001, ISBN-10 3540419535.

[Cha86] Charpentier, F.: Pitch detection using the short-term phase spectrum. Proc. ICASSP 1986, Tokyo, Japan, 1986, S. 113 – 116.

[CHIL] CHIL – Computers in the Human Interaction Loop. http://chil.server.de/servlet/is/101/.

[CMU+95] Cole, R.; Mariani, J.; Uszkoreit, H.; Zaenen, A. und Zue, V.: Survey of the State of the Art in Human Language Technology. Center for Spoken Language Understanding CSLU, Carnegie Mellon University, Pittsburgh, USA, 1995.

[CW02] Chu, W. T. und Warnock, A. C. C.: Detailed directivity of sound fields around human talkers. Technical Report RR-104, National Research Council Canada, 2002.

[dCK01] de Cheveigné, A. und Kawahara, H.: Comparative evaluation of F0 estimation algorithms. Proc. EUROSPEECH 2001, Aalborg, Dänemark, 2001, S. 2451 – 2454.

[dCK02] de Cheveigné, A. und Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustic Society of America, Vol. 111(2002), No. 4, S. 1917 – 1930.

[DF39] Dunn, H. K. und Farnsworth, D. W.: Exploration of pressure field around the human head during speech. Journal of the Acoustical Society of America, Vol. 10(1939), No. 1, S. 184 – 199.

[DFP94a] Drullman, R.; Festen, J. M. und Plomp, R.: Effect of temporal envelope smearing on speech perception. Journal of the Acoustic Society of America, Vol. 95(1994), No. 2, S. 1053 – 1064.

[DFP94b] Drullman, R.; Festen, J. M. und Plomp, R.: Effect of reducing slow temporal modulations on speech perception. Journal of the Acoustic Society of America, 95(1994), No. 5, S. 2670 – 2680.

[DHM05] Delcroix, M., Hikichi, T. und Miyoshi, M.: Blind dereverberation algorithm for speech signals based on multi-channel linear prediction. Acoustical Science and Technology, Vol. 26(2005), No. 5, S. 432 – 439.

[DHM07] Delcroix, M., Hikichi, T. und Miyoshi, M.: Precise dereverberation using multichannel linear prediction. IEEE Transactions on Audio, Speech and Language Processessing, Vol. 15(2007), No. 2, S. 430 – 440.

[Dic99] Dickreiter, M.: Mikrofon Aufnahmetchnik. Auflage: 3., völlig neu bearbeitete Auflage, S. Hirzel Verlag, 2003, ISBN 3777611999.

[DIN03a] Deutsches Institut für Normung: DIN 1343:1990-01 Referenzzustand, Normzustand, Normvolumen; Begriffe und Werte. Beuth Verlag, 2003.

[DIN03b] Deutsches Institut für Normung: DIN EN 60268-16 Objektive Bewertung der Sprachverständlichkeit durch den Sprachübertragungsindex (IEC 60268-16:2003); Deutsche Fassung EN 60268-16:2003. In DIN EN 60268: Elektroakustische Geräte, Teil 16, Beuth Verlag, 2003.

[DIN72] Deutsches Institut für Normung: DIN 45630-1 Grundlagen der Schallmessung; Physikalische und subjektive Größen von Schall. 1972.

[Dro08] Droppo, J. und Acero, A.: Environmental Robustness. In: Springer Handbook of Speech Processing. Benesty, J.; Sondhi, M. M.; Huang, Y. (Eds.), XXXVI, Springer-Verlag New York, 2008, ISBN 978-3-540-49125-5, S. 653 – 679.

[Duc06] Duckhorn, F.: Complex Harmonic Filter Analysis (CHFA) – Eigenschaften und Eignung zur Periodenmarkierung von Sprachsignalen. Studienarbeit, Institut für Akustik und Sprachkommunikation, Technischen Universität Dresden, 2006.

[Dud39] Dudley, H.: Remaking speech. Journal of the Acoustic Society of America, Vol. 11(1939), No. 2, S. 169 – 177.

[ECESS] European Center of Excellence in Speech Synthesis. www.ecess.eu.

[Eic07] Eichner, M.: Sprachsynthese und Spracherkennung mit gemeinsamen Datenbasen, Akustische Analyse und Modellierung. Dissertation, Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, 2007.

[EKPH06] Estelmann, J.; Koloska, U.; Petrick, R. und Hirschfeld, D.: VAD-Verfahren. In: Studientexte zur Sprachkommunikation, Vol. 31, Tagungsband der 17. Konferenz für Elektronische Sprachsignalverarbeitung (ESSV), Freiberg 2006, ISBN 3-938863-74-9, S. 221 – 226.

[EM07] Eneman, K. und Moonen, M.: Multimicrophone Speech Dereverberation: Experimental Validation. EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2007, Article ID 51831.

[ETS03a] ETSI Standard ETSI ES 201 108 V1.1.3: Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms. European Telecommunications Standards Institute, 2003.

[ETS03b] ETSI Standard ETSI ES 202 211 V1.1.1: Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm. European Telecommunications Standards Institute, 2003.

[ETS05a] ETSI Standard ETSI ES 202 050 V1.1.3: Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. European Telecommunications Standards Institute, 2005.

[ETS05b] ETSI Standard ETSI ES 202 212 V1.1.2: Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm. European Telecommunications Standards Institute, 2005.

[EWH00] Eichner, M.; Wolff, M. und Hoffmann, R.: A Unified Approach for Speech Synthesis and Speech Recognition Using Stochastic Markov Graphs. Proc. ICSLP 2000, Beijing, China, 2000, S. 701 – 704.

[Fan60] Fant, G.: Acoustic theory of speech production. Mouton, The Hague, 1960.

[Feh06] Fehér, T.: Entwurf und Optimierung eines Richtmikrofons – bestehend aus mehreren Einzelmikrofonen in Verbindung mit digitaler Signalverarbeitung. Diplomarbeit, Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, 2006.

[Fla60] Flanagan, J. L.: Analog measurements of sound radiation from the mouth. Journal of the Acoustical Society of America, Vol. 32(1960), No. 12, S. 1613 – 1620.

[Fur86] Furui, S.: Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 32(1984), No. 1, S. 52 – 59.

[Fur01] Furuya, K.: Noise reduction and dereverberation using correlation matrix based on the multiple-input/output inverse-filtering theorem (MINT). Proc. International Workshop on Hands-free Speech Communication, Japan, 2001, S. 59 – 62.

[Gal95] Gales, M. J.: Model based techniques for noise robust speech recognition. Ph.D. Thesis, Cambridge University, 1995.

[Gar07] Gareta, A. A.: A multi-microphone approach to speech processing in a smart-room environment. Ph.D. Thesis, Universitat Politècnica de Catalunya, Barcelona, Spanien, 2007.

[GB99] Griebel, S. und Brandstein, M.: Wavelet tranform extrema clustering for multi-channel speech deverberation. Proc. WASPAA 1999, New Paltz, USA, 1999.

[Ger03] Gerhard, D.: Pitch Extraction and Fundamental Frequency: History and Current Techniques, Technical Report, Dept. of Computer Science, University of Regina, 2003.

[GHE96] Greenberg, S.; Hollenback, J. und Ellis, D.: Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. Proc ICSLP 96, Philadelphia, USA, 1996, S. 32 – 35.

[GMF01] Gillespie, B. W.; Malvar, H. S. und Florencio, D. A.: Speech dereverberation via maximum-kurtosis subband adaptive filtering. Proc. ICASSP 2001, Salt Lake City, Utah, USA, 2001, S. 3701 – 3704.

[GMOS99] Giuliani, D.; Matassoni, M.; Omologo, M. und Svaizer, P.: Training of HMM with filtered speech material for hands-free recognition. Proc. ICASSP 1999, Phoenix, USA, 1999, S. 449 – 452.

[GNW03] Gaubitch, N. D.; Naylor, P. A. und Ward, D. B.: On the use of linear prediction for dereverberation of speech. Proc. IWAENC 2003, Kyoto, Japan, 2003, S. 99 – 102.

[GNW04] Gaubitch, N. D.; Naylor, P. A. und Ward, D. B.: Multi-microphone speech dereverberation using spatio-temporal averaging. Proc. EUSIPCO 2004, Wien, Österreich, 2004, S. 809 – 812.

[Gon95] Gong, Y.: Speech recognition in noisy environments: A survey. Speech communication, Vol. 16(1995), No. 3, S. 261 – 291.

[Gre96] Greenberg, S.: Understanding speech understanding – towards a unified theory of speech perception. Proc. of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, W. A. Ainsworth und S. Greenberg (Eds.), Keele, England, 1996, S. 1 – 8.

[Gre97] Greenberg, S.: On the origins of speech intelligibility in the real world. Proc. ESCA-NATO Tutorial and Research Workshop on Robust speech recognition for unknown communication channels, Pont-a-Mousson, Frankreich, 1997.

[Gre99] Greenberg, S.: Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. Speech Communication, Vol. 29(1999), No. 2 – 4, S. 159 – 176.

[Gre05] Greß, O.: Untersuchung des Einflusses von Nachhall auf die Automatische Spracherkennung. Bachelorarbeit, Lehrstuhl für Multimediakommunikation und Signalverarbeitung, Universität Erlangen-Nürnberg, 2005.

[Gre07] Greenberg, S. The History and Biology of the Modulation Spectrum. Tutorial at the INTERSPEECH 2007. Antwerp, Belgien, 2007.

[Gru04] Gruber, C.: Entwicklung und Implementierung eines einsatzfähigen Algorithmus zur Verbesserung von Freisprechanwendungen im Kfz. Studienarbeit, Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, 2004.

[Haa51] Haas, H.: Über den Einfluss eines Einfachechos auf die Hörsamkeit von Sprache. Acustica, Vol. 1(1951), S. 49-58.

[HAA+08] Hoffmann, R.; Alisch, L.-M.; Altmann, A.; Fehér, T.; Petrick, R.; Wittenberg, S. und Hermkes, R.: The Acoustic Front-end in Scenarios of Interaction Research. In: Esposito, A., et al. (Eds.): Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction. Selected papers from COST Action 2102 International Workshop. Berlin etc.: Springer-Verlag, 2008.

[Hab05] Habets, E. A. P.: Multi-Channel Speech Dereverberation based on a Statistical Model of Late Reverberation. Proc. ICASSP 2005, Philadelphia, USA, 2005, S. 173 – 176.

[Hab06] Habets, E.: Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement. Ph.D. Thesis, Technische Universiteit Eindhoven, Niederlande, 2007.

[Hau98] Haulick, T.: Robuste Geräuschreduktion für Kompakte Mikrofonanordnungen. Dissertation, Universität Karlsruhe, 1998.

[HBC08] Huang, Y.; Benesty, J. und Chen, J.: Dereverberation. In: Benesty, J; Sondhi, M. M.; Huang, Y. (Eds.): Springer Handbook of Speech Processing. 2008, XXXVI, Springer-Verlag New York, 2008, ISBN 978-3-540-49125-5, S. 929 – 943.

[HBK+02] Hirschfeld, D.; Bechstein, J.; Koloska, U.; Richter, T. und Petrick, R.: Entwicklungsschritte eines Hardware-Kommandoworterkenners mit minimalem Footprint. In: Studientexte zur Sprachkommunikation, Vol. 24, Tagungsband der 13. Konferenz für Elektronische Sprachsignalverarbeitung (ESSV), w.e.b. Universitätsverlag, Dresden, 2002, ISBN 3-935712-72-3, S. 182 – 189.

[HDM06] Hikichi, T.; Delcroix, M. und Miyoshi, M.: On robust inverse filter design for room transfer function fluctuations. Proc. EUSIPCO 2006, Florence, Italien, 2006.

[HDM07] Hikichi, T.; Delcroix, M. und Miyoshi, M.: Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations. EURASIP Journal on Advances in Signal Processing, Vol. 2007, Article ID 34013.

[Her88] Hermes, D. J.: Measurement of pitch by subharmonic summation. Journal of the Acoustic Society of America, Vol. 38(1988), No. 1, S. 257 – 264.

[Her97] Hermansky, H.: The Modulation Spectrum in the Automatic Recognition of Speech. Proc. ASRU, IEEE Signal Processing Society, Editors Furui, S.; Juang, B.-H. und W, Chou, 1997, S. 140 – 147.

[Hes83] Hess, W. J.: Pitch Determination of Speech Signals. Springer-Verlag, New York, 1983, ISBN 0387119337.

[Hes92] Hess, W. J.: Pitch and Voicing Determination. In: Furui, S. und Sondhi, M. M. (Eds.): Advances in Speech Signal Processing. Marcel Dekker Inc. New York, 1992, ISBN 0824785401, S. 3 – 48.

[Hes08] Hess, W. J.: Pitch and Voicing Determination of Speech with an Extention Towards Music Signals. In: Springer Handbook of Speech Processing. Benesty, J; Sondhi, M. M.; Huang, Y. (Eds.), XXXVI, Springer-Verlag New York, 2008, ISBN 978-3-540-49125-5, S. 181 – 211.

[HEW07] Hoffmann, R.; Eichner, M.; Wolff, M.: Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system. In: Esposito, A.; Faundez-Zanuy, M.; Keller, E. und Marinaro, M. (Eds.): Verbal and Nonverbal Communication Behaviours. Springer-Verlag Berlin 2007, ISBN 978-3-540-76441-0, S. 200 – 218.

[HF06] Hirsch, H.-G. und Finster, H.: A New HMM Adaptation Approach for the Case of a Hands-free Speech Input in Reverberant Rooms. Proc. INTERSPEECH 2006, Pittsburgh, USA, 2006, S. 781 – 784.

[Hir88] Hirsch, H. G.: Automatic Speech Recognition in Rooms. In: Lacourne, Chehikian, Martin und Malbos (Eds.): Signal Processing IV: Theories and Applications. Elsevier Science Publishers B. V. (North Holland), EURASIP, 1988, S. 1177 – 1180.

[HJ07] Hussein, H. und Jokisch, O.: Hybrid electroglottograph and speech signal based algorithm for pitch marking. Proc. INTERSPEECH 2007, Antwerp, Belgien, S. 1653 – 1656.

[HKKP06] Höge, H.; Kotnik, B.; Kacic, Z.; Pfitzinger, H. R.: Evaluation of Pitch Marking Algorithms. Proc. ITG-Fachtagung Sprachkommunikation 2006, Kiel, Germany, 2006.

[HM94a] Heckl, M. und Müller, H.A.: Taschenbuch der technischen Akustik. 2. Aufl. Springer-Verlag, 1994, ISBN 3-540-54473-9.

[HM94b] Hermansky, H. und Morgan, N.: RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, Vol. 2(1994), No. 4, S. 578 – 589.

[HNH+05] Haderlein, T.; Nöth, E.; Herbordt, W.; Kellermann, W. und Niemann, H.: Using Arti-

ficially Reverberated Training Data in Distant-Talking ASR. In: Matoušek, V.; Mautner, P. und Pavelka, T. (Eds.): Text, Speech and Dialogue 2005, Springer-Verlag Berlin Heidelberg, 2005, ISBN 978-3-540-28789-6, S. 226 – 233.

[HNKT98] Hirobayashi, S.; Nomura, H.; Koike, T.; und Tohyama, M.: Speech waveform recovery from a reverberant speech signal using inverse filtering of the power envelope transfer function (in Japanisch). IEICE Transactions on Acoustics, Vol. J81-A(1996), S. 1323 – 1330.

[Hoe07] Höge, H.: Basic Parameters in Speech Processing The Need for Evaluation. Archives of Acoustics, Vol. 32(2007), No. 1, Warszawa, Polen, 2007, S. 67 – 74.

[Hof98] Hoffmann, R.: Signalanalyse und -erkennung. Springer-Verlag, Berlin Heidelberg New York, 1998, ISBN 3-540-63443-6.

[Hon08] Honda, K.: Physiological Process of Pitch Production. In: Springer Handbook of Speech Processing. Benesty, J; Sondhi, M. M.; Huang, Y. (Eds.), XXXVI, Springer-Verlag New York, 2008, ISBN 978-3-540-49125-5, S. 7 – 26.

[HS73] Houtgast, T. und Steeneken, H. J. M.: The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acustica, Vol. 28(1973), S. 66 – 73.

[HS80] Houtgast, T. und Steeneken, H. J. M.: A physical method for measuring speech transmission quality. Journal of the Acoustical Society of America, Vol. 67(1980), No. 1, S. 318 – 326.

[HS84] Houtgast, T. und Steeneken, H. J. M.: A multi-language evaluation of the RASTI-method for estimation speech intelligibility in auditoria. Acustica 54(1984), S. 185 – 199.

[HS85] Houtgast, T. und Steeneken, H. J. M.: A review of the MTF concept in room acoustics and its use for estimation speech intelligibility in auditoria. The Journal of the Acoustical Society of America, Vol. 77(1985), No. 3, S. 1069 – 1077.

[HSP80] Houtgast, T.; Steeneken, H. J. M. und Plomb, R.: Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. Acustica, Vol. 46(1980), S. 60 – 72.

[HTK02] The HTK Book. V3.2., Engineering Department at the Cambridge University, 2002.

[HV04] Halkosaari, T. und Vaalgamaa, M.: Directivity of human and artificial speech. Joint Baltic-Nordic Acoustics Meeting 2004, Mariehamn, Åland, 2004.

[ISO97] ISO 3382, Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters. International Standardisation Organisation. Reference Number IS0 3382: 1997(E).

[ITU96] ITU-T Recommendation P.800: Methods for subjective determination of transmission quality. Genf, Schweiz, 1996.

[ITU01] ITU-T Recommendation P.862: Perceptial evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunications Union (ITU-T), 2001.

[IUA01] Ishimoto, Y.; Unoki, M.; Akagi, M.: A fundamental frequency estimation method for noisy speech based on instantaneous amplitude and frequency. Proc. EUROSPEECH 2001, Aalborg, Dänemark, 2001, S. 2439 – 2442.

[Jan68] Januška, I.: Experimentally stated correlation between objective echogramms evaluation and speech intelligibility. Archivum Akustiki, 1968, S. 140.

[JH95] Junqua, J.-C. und Haton, J.-P.: Robustness in Automatic Speech Recognition, Fundamentals and Applications. Kluwer Academic Publisher, Boston Dordrecht London, 1995, ISBN 0792396464.

[Jos02] Josifovski, L.: Robust Automatic Speech Recognition with Missing and Unreliable Data. Dissertation, Department of Computer Science, University of Sheffield, UK, 2002.

[KAHP99] Kanedera, N.; Arai, A.; Hermansky, H. und Pavel, M.: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication, Vol. 28(1999), No. 1, S. 43 – 55.

[KdCP98] Kawahara, H.; de Cheveigné, A.; Patterson, R. D.: An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite. Proc. ICSLP 1998, Sydney, Australien, 1998.

[KDNM07] Kinoshita, K., Delcroix, M., Nakatani, T. und Miyoshi, M.: Multi-step linear prediction based speech dereverberation in noisy reverberant environment. Proc. INTERSPEECH 2007, Antwerp, Belgien, 2007, S. 3 – 15.

[KHK06] Kotnik, B.; Höge, H.; Kacic, Z.: Evaluation of Pitch Detection Algorithms in Adverse Conditions. Proc. Speech Prosody 2006, Dresden, Deutschland, 2006.

[KKdCP99] Kawahara, H.; Katayose, H.; de Cheveigné, A. und Patterson, R. D.: Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. Proc. EUROSPEECH 1999, Budapest, Ungarn, 1999, S. 2781 – 2784.

[Kle02] Kleinschmidt, M: Methods for capturing spectro temporal modulations in automatic speech recognition. Acustica united with Acta Acustica, Vol. 88(2002), No. 3, S. 416 – 422.

[KM04] Kuttruff, H. und Mommerz, E.: Raumaustik. In: Müller, G. und Möser, M. (Hrsg.): Taschenbuch der Technischen Akustik. Springer-Verlag Berlin Heidelberg New York, 2004, ISBN 3-540-41242-5, S. 331 – 366.

[KNM05] Kinoshita, K.; Nakatani, T. und Miyoshi, M.: Fast estimation of a precise dereverberation filter based on speech harmonicity. Proc. ICASSP 2005, Philadelphia, USA, 2005, S. 1073 – 1076.

[KNM06] Kinoshita, K.; Nakatani, T. und Miyoshi, M.: Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation. Proc. ICASSP 2006, Toulouse, Frankreich, 2006, S. 817 – 820.

[KR99] Kennedy, R. und Radlović, B.: Iterative cepstrum-based approach for speech dereverberation. Proc. of the fifth International Symposium on Signal Processing and its Applications – ISSPA 1999, 1999, S. 55 – 58.

[KSS97] Kunieda, N.; Shimamura, T. und Suzuki, J.: Pitch extraction by using autocorrelation function on the log spectrum. IEICE Transactions on Acoustics. Vol. J80-A(1997), No. 3, S. 435 – 443.

[Kuc04] Kuchling, H.: Taschenbuch der Physik. 18. Auflage, Fachbuchverlag Leipzig, 2004, ISBN 3-446-22883-7.

[Kue69] Kürer, R.: Gewinnung von Einzelkriterien bei der Impulsmessung in der Raumakustik. Acustica 21(1969), S. 370 – 372.

[Kut00] Kuttruff, H.: Room Acoustics. Fourth edition, Spon Press, 2000, ISBN 978-0419245803.

[Kut04] Kuttruff, H.: Akustik. Hirzel-Verlag Stuttgart, 2004, ISBN 3-7776-1244-8.

[Lad95] Ladefoged, P.: Elements of Acoustic Phonetics. University Of Chicago Press, First Edition, 1995, ISBN-10 0226467643, ISBN-13 978-0226467641.

[LB61] Lochner, J. P. A. und Burger, J. F.: The intelligibility of speech under reverberant conditions. Acustica 11(1961), S. 195.

[LBD01] Lebart, K.; Boucher, J. M. und Denbigh, P. N.: A new method based on spectral subtraction for speech dereverberation. Acta Acoustica, Vol. 87(2001), No. 3, S. 359 – 366.

[Leh74] Lehmann, U.: Untersuchung zur Bestimmung des Raumeindrucks bei Musikdarbietungen und Grundlagen der Optimierung. Dissertation, Technische Universität Dresden, 1974.

[LMP87] Lippmann, R. P.; Martin, E. A. und Paul, D. P.: Multistyle training for robust isolated-word speech recognition. Proc. ICASSP 1987, Dallas, USA, 1987, S. 709 – 712.

[Loh07] Lohde, K.: Erhöhung der Robustheit eines Spracherkenners gegen Raumeinflüsse. Diplomarbeit, Fachbereich Elektrotechnik, Hochschule für Technik und Wirtschaft Dresden (FH), 2007.

[Lor08] Lorenz, M.: Maßnahmen zur Steigerung der Erkennungsleistung von Spracherkennern bei verhallter Sprache. Studienarbeit, Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, 2008.

[LS82] Langhans, T. und Strube, H. W.: Speech enhancement by nonlinear multiband envelope filtering. Proc. ICASSP 1982, Paris, Frankreich, 1982, S. 156 – 159.

[LSN+07] Luengo, I., Saratxaga, I., Navas, E., Hernaez, I., Sanchez, J., Sainz, I.: Evaluation of pitch detection algorithms under real conditions. Proc. ICASSP 2007, Honolulu, Hawaii, USA, 2007, S. 1057 – 1060.

[LUA06] Lu, X.; Unoki, M. und Akagi, M.: A robust feature extraction based on the MTF concept for speech recognition in reverberant environment. Proc. INTERSPEECH 2006, Pittsburgh, 2006, S. 2546 – 2549.

[LUA07] Lu, X.; Unoki, M. und Akagi, M.: Comparative evaluation of MTF-based feature extraction for speech recognition in reverberant environments. Proc. SPECOM 2007, Moscow, Russia, 2007, S. 124–133.

[LUA08] Lu, X.; Unoki, M. und Akagi, M.: Comparative evaluation of modulation-transfer-function-based blind restoration of sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems. Acoustic Science & Technology, Vol. 29(2008), No. 6, S. 351 – 361.

[Mar72] Markel, J.: The SIFT algorithm for fundamental frequency estimation. IEEE Transactions on Audio and Elektroacoustics, Vol. AU-20(1972), No. 5, S. 367 – 377.

[Mar82] Martin, P.: Comparison of Pitch detection by cepstrum and spectral comb analysis. Proc. ICASSP 1982, Tokyo, Japan, S. 180 – 183.

[McK86] McKendree, F. S.; Directivity indices of human talkers in English speech. Proc. International Conference on Noise Control Engineering 1986. Noise Control Foundation, Cambridge, USA, 1986, S. 911 – 916.

[Mei98] Meinschaefer, J.: Silbe und Sonorität in Sprache und Gehirn. Dissertation, Fakultät für Philologie, Ruhr-Universität Bochum, 1998.

[MH83] Mourjopoulos J. und Hammond, J. K.: Modelling and enhancement of reverberant speech using an envelope convolution method. Proc. ICASSP 1983, Boston, USA, 1983, S. 1144 – 1147.

[MHK+03] Maase, J.; Hirschfeld, D.; Koloska, U.; Westfeld,T. und Helbig, J.: Towards an Evaluation Standard for Speech Control Concepts in Real-World Scenarios. Proc. EUROSPEECH 2003, Geneva, Schweiz, 2003, S. 1553 – 1556.

[MHMH07] Molla, K.I.; Hirose, K.; Minematsu, N. und Hasan, K.: Voiced/Unvoiced Detection of Speech Signals Using Empirical Mode Decomposition Model. International Conference on Information and Communication Technology – ICICT 2007, Dhaka, Bangladesh,

2007, S. 311 – 314.

[MK88] Miyoshi, M. und Kaneda, Y.: Inverse filtering of room acoustics. IEEE Transactions on Acoustics, Speech Signal Processing, Vol. 36(1988), No. 2, S. 145 – 152.

[MM04] Müller, G. und Möser, M. (Hrsg.): Taschenbuch der Technischen Akustik. Springer-Verlag Berlin Heidelberg New York, 2004, ISBN 3-540-41242-5, S. 331 – 366.

[MMN+02] Macho, D.; Mauuary, L.; Noé, B.; Cheng, Y. M.; Ealey, D.; Jouvet, D.; Kelleher, H.; Pearce, D. und Saadoun, F.: Evaluation of a noise-robust DSR front-end on Aurora databases. Proc. ICSLP 2002, Denver, USA, 2002, S. 17 – 20.

[Moe01] Möbius, B.: German and Multilingual Speech Synthesis. Habilitationsschrift, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, Universität Stuttgart, AIMS Vol. 7, No. 4, 2001.

[Moo74] Moorer, J. A.: The optimum comb method of pitch period analysis of continious digitized speech. IEEE Transactions in Acoustics, Speech and Signal Processing, Vol. 22(1974), No. 5, S. 330 – 338.

[MP78] Moreno, A. und Pfretzschner, J.: Human Head Directivity in Speech Emission: A New Approach. Acoustics Letters, Vol. 1(1978), S. 78 – 84.

[MRS96] Moreno, P. J.; Raj, B. und Stern, R. M.: A vector Taylor series approach for environment-independent speech recognition. Proc. ICASSP 1996, Atlanta, USA, 1996, S. 733 – 736.

[NA79] Neely, S. T. und Allen, J. B.: Invertibility of a room impulse response. Journal of the Acoustic Society of America, Vol. 66(1979), No. 1, S. 165 – 169.

[NG05] Naylor, P. A. und Gaubitch, N. D.: Speech dereverberation, Proc. IWAENC 2005, Eindhoven, Niederlande, 2005, S. 121 – 124.

[NGMH07] Neumann, J.; Gasas, J. R.; Macho, D.; Hidalgo, J. R.: Integration of audio-visual sensors and technologies in a smart room. In: Personal and Ubiquitous Computing, Springer-Verlag London, ISSN 1617-4909 (print) 1617-4917 (online), 2007.

[Nie56] Niese, H.: Vorschlag für die Definition und Messung der Deutlichkeit nach subjektiven Grundlagen. Hochfrequenztechnik und Elektroakustik, Vol. 65(1956), Heft 1, S. 4.

[NJY+08] Nakatani, T. Juang, B.-H.; Yoshioka, T.; Kinoshita, K.; Delcroix, M. und Miyoshi, M.: Speech Dereverberation Based on Maximum-Likelihood Estimation With Time-Varying Gaussian Source Model. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16(2008), No. 8, S. 1512 – 1527.

[NKM07] Nakatani, T.; Kinoshita, K. und Mihyoshi, M.: Harmonicity based blind dereverberation for single-channel speech signals. IEEE Transactians of Audio, Speech and Language Processing, Vol. 15(2007), No. 1, S. 80 – 95.

[NM03] Nakatani, T. und Miyoshi, M.: Blind dereverberation of single channel speech signal based on harmonic structure, Proc. ICASSP 2003, Hong Kong, 2003, S. 92 – 95.

[NMK05] Nakatani, T.; Mihyoshi, M. und Kinoshita, K.: Single-Microphone blind dereverberation. In: Benesty, J.; Makino, S.; Chen, J. (Eds.): Speech Enhancement. Springer-Verlag Berlin Heidelberg, 2005, Chapter 11, S. 247 – 270.

[Nol64] Noll, A. M.: Short-time spectrum and "cepstrum" techniques for vocal-pitch detection. Journal of the Acoustic Society of America, Vol. 36(1964), No. 2, S. 226 – 302.

[Nol66] Noll, A. M.: Cepstrum pitch determination. Journal of the Acoustic Society of America, Vol. 41(1966), No. 2, S. 293 – 309.

[Oha99] Ohara, Y.: Performing gender through voice pitch: a cross-cultural analysis of Japanese and American English. In: Pasero., U. et al.: Perceiving and Performing Gender,

Westdeutscher Verlag, Wiesbaden, 1999, ISBN 3-531-13379-9, S. 105 – 116.

[OMSG97] Omologo, M.; Matassoni, M.; Svaizer, P. und Giuliani, D.: Microphone array based speech recognition with different talker-array positions. Proc. ICASSP 1997, München, Deutschland, 1997, S. 227 – 230.

[OS89] Oppenheim, A. V. und Schäfer, R. W.: Discrete-Time Signal Processing. Prentice-Hall, USA, 1989, ISBN-10: 013216292X.

[OS04] Oppenheim, A. V. und Schäfer, R. W.: Zeitdiskrete Signalverarbeitung. 2. überarbeitete Auflage, Oldenbourg-Verlag, München, 2004, ISBN-10: 3827370779.

[Pet01] Petrick, R: Maßnahmen zur Erhöhung der Robustheit eines schnellen Kommandoworterkenners. Diplomarbeit, Fachbereich Elektrotechnik, Hochschule für Technik und Wirtschaft-Dresden (FH), 2001.

[PGF04] Petrick, R.; Gruber, C. und Fenske, M.: Ein effektiver Algorithmus zur kombinierten Echounterdrückung und Geräuschreduktion in Freisprechanwendungen. In: Studientexte zur Sprachkommunikation, Vol. 30, Tagungsband der 15. Konferenz für Elektronische Sprachsignalverarbeitung (ESSV), w.e.b. Universitätsverlag, Cottbus, 2004, ISBN 3-937672-65-5, S. 236 – 243.

[PHGK05] Petrick, R.; Hirschfeld, D.; Gruber, C. und Kienast, G.: Comparison of Signal Enhancement Techniques in Communications and Speech Control Tasks for a Single-DSP in-Car Application. Biennial on DSP for in-Vehicle and Mobile Systems, Sesimbra, Portugal, 2005, Paper M2-6.

[PHHJ03] Petrick, R.; Hirschfeld, D.; Hoffmann, R. und Jokisch, O.: Verbkey – a DSP based Speech Control for the Automotive Environment. Workshop on DSP in Mobile and Vehicular Systems, Nagoya, Japan, 2003.

[PHHJ04] Petrick, R.; Hirschfeld, D.; Hoffmann, R. und Jokisch, O.: Verbkey – a DSP based Speech Control for the Automotive Environment. In: Abut, H.; Hansen, J. H. L.; Takeda, K. (Eds.): DSP for In-Vehicle and Mobile Systems. Chapter 12, Springer-Verlag, 2005, ISBN-10: 0387229787, S. 179 – 192.

[PJH07] Petrick, R.; Jokisch, O. und Hoffmann, R.: The Infuence of Reverberation: Speech Recognition versus Human Perception. Proc. SPECOM 2007, Moscow, Russland, 2007.

[PKH05] Petrick, R.; Kinast, G. und Hirschfeld, D.: Influence of a Single Channel and a Multi Channel Noise Reduction on the Recognition of Noisy Speech. In: Studientexte zur Sprachkommunikation, Vol. 36, Proceedings of the 16th Conference on Electronic Speech Signal Processing (ESSP), w.e.b. Universitätsverlag, Praha, Tschechien, 2005, ISBN 3-938863-17-X, S. 159 – 166.

[PLLH08] Petrick, R.; Lohde, K.; Lorenz, M. und Hoffmann, R.: A new feature analysis method for robust ASR in reverberant environments based on the harmonic structure of speech. Proc. EUSIPCO 2008, Lausanne, Schweiz, 2008.

[PLU+08a] Petrick, R.; Lu, X.; Unoki, M.; Akagi, M.; Hoffmann, R.: Robust Front End Processing for Speech Recognition in Reverberant Environments: Utilization of Speech Characteristics. Proc. INTERSPEECH 2008, Brisbane, Australien, 2008, S. 658 – 661.

[PLU+08b] Petrick, R.; Lu, X.; Unoki, M.; Akagi, M.; Hoffmann, R.: Robust Front End Processing for Speech Recognition in Reverberant Environments: Utilization of Speech Properties. Technical Report of IEICE, Morioka, Iwate, Japan, 2008, S. 7 – 12.

[PLWH07] Petrick, R.; Lohde, K.; Wolff, M. und Hoffmann, R.: The harming part of room acoustics for automatic speech recognition. Proc. of INTERSPEECH 2007, Antwerp, Belgien, 2007, S. 1094 – 1097.

[Pom95] Pompino-Marschall, B.: Einführung in die Phonetik. de Gruyter, Berlin New York,

1995, ISBN 3-11-014763-7.

[PS94] Petropulu, A. und Subramaniam, S.: Cepstrum based deconvolution for speech dereverberation. Proc. ICASSP 1994, Adelaide, Australien, 1994, S. 9 – 12.

[PUM+08] Petrick, R.; Unoki, M.; Mittal, A.; Segura, C. und Hoffmann, R.: A comprehensive study on the effects of room reverberation on fundamental frequency estimation. Proc. INTERSPEECH 2008, Brisbane, Australien, 2008, S. 131 – 134.

[QBC88] Quackenbush, S. R.; Barnwell, T. P. und Clements, M. A.: Objective Measures of Speech Quality. Englewood Cliffs New Jersey: Prentice-Hall, Inc., 1988, ISBN 0136290566.

[QSS02] Quast, H.; Schreiner, O. und Schroeder, M. R.: Robust pitch tracking in the car environment. Proc. ICASSP 2002, Orlando, USA, 2002, S. 353 – 356.

[Rab89] Rabiner, L. R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(1989), No. 2, S. 257-285.

[RAS75] Reichardt, W.; Abdel Alim, O. und Schmidt, W.: Definition und Meßgrundlage eines objektiven Maßes zur Ermittlung und Grenze zwischen brauchbarer und unbrauchbarer Durchsichtigkeit bei Musikdarbietungen. Acustica, Vol. 32(1975), S. 126.

[RCRM76] Rabiner, L. R.; Cheng, M. J.; Rosenberg, A. E.; McGonegal, C. A.: A comparative performance study of several pitch detection algorithms. IEEE Transactions on Audio, Signal, and Speech Processing, Vol. 24(1976), No. 5, S. 399 – 418.

[Rei75] Walter Reichardt et al.: Zusammenhang zwischen Klarheitsmaß C und anderen raumakustischen Kriterien. Zeitschrift für elektrische Informations- und Energietechnik, Vol. 5(1975), S. 144.

[RJ93] Rabiner, L. R. und Juang, B.-H.: Fundamentals of Speech Recognition. Prentice Hall, Engelwood Cliffs, 1993, ISBN 0-13-015157-2.

[RLM97] Rouat, J.; Liu, Y. C. und Morissette, D.: A pitch determination and voiced/unvoiced decision algorithm for noisy speech. Speech Communication, Vol. 21(1997), No. 3, S. 191 – 207.

[RNS06] Raut, C. K.; Nishimoto, T. und Sagayama, S.: Model Adaptation for Long Convolutional Distortion by Maximum Likelihood State Filtering Approach, Proc. ICASSP 2006, Toulouse, Frankreich, 2006, S. 1133 – 1137.

[RPH03] Richter, T.; Petrick, R. und Hirschfeld, D.: Robuste Phrasendetektion durch zweistufige Sprach/Pause-Detektion. In: Studientexte zur Sprachkommunikation, Vol. 28, Tagungsband der 14. Konf. für Elektron. Sprachsignalverarbeitung (ESSV), w.e.b. Universitätsverlag, Karlsruhe, 2003, ISBN 3-935712-83-9, S. 46 – 53.

[RS66] Reichardt, W. und Schmidt, W.: Die hörbaren Stufen des Raumeindrucks bei Musik. Acustica, Vol. 17(1966), S. 175.

[RSC+74] Ross, M. J.; Shaffer, H. L.; Cohen, A.; Freudberg, R. und Manley, H. J.: Average magnitude difference function pitch extractor. IEEE Transactions in Acoustics, Speech and Signal Processing, Vol. 22(1974), No. 5, S. 353 – 361.

[RSLA74] Reichardt, W.; Schmidt, W.; Lehmann, U. und Ahnert, W.: Definition und Messgrundlagen eines "wirksamen Hallabstandes" als Maß für den Raumeindruck bei Musikdarbietungen. Zeitschrift für elektrische Informations- und Energietechnik, Vol. 4(1974), S. 225 – 233.

[RSS75] Rabiner, L.; Sambur, M. und Schmidt, C.: Applications of a nonlinear smoothing algorithm to speech processing. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 23(1975), No. 6, S. 552 – 557.

[Rud99] Rudolph, T.: Evolutionäre Optimierung Schneller Worterkenner. Dissertation, Insti-

tut für Akustik und Sprachkommunikation, Technische Universität Dresden, Dresden, 1999.

[Sag78] Sagayama, S. et al.: Pitch extraction using a lagwindow method (in Japanisch). Proc. IECEJ Meeting, 1978.

[SB68] Sobolev, V. N.; Baronin S. P.: Investigation of the shift Method for Pitch Determination (in Russisch). Elektrosvyaz, Vol. 12, 1968, S. 30 – 36.

[SBMK93] Smith, C. L.; Browman, C. P.; McGowan, R. S. und Kay, B.: Extracting dynamic parameters from speech movement data. Journal of the Acoustic Society of America, Vol. 93(1993), No. 3, S. 1580 – 1588.

[SC99] Shire, M. L. und Chen, B. Y.: Data-driven RASTA filters in reverberation. Proc. EUROSPEECH 1999, Genova, Italien, 1999, S. 1123 – 1126.

[SC00] Shire, M. L. und Chen, B. Y.: Data-driven RASTA filters in reverberation. Proc. ICASSP 2000, Istanbul, Türkei, 2000, S. 1627 – 1630.

[Sch65] Schroeder, M. R.: New Method of Measuring Reverberation Time. Journal of the Acoustic Society of America, Vol. 37(1965), No. 3, S. 409 – 412.

[Sch78] Schroeder, M. R.: Modulation transfer functions: definition and measurement. Abstract in IEEE Transactions on Acoustics Speech and Signal Processing, Vol. 26(1978), S. 180.

[Sch79] Schmidt, W.: Raumakustische Gütekriterien und ihre objektive Bestimmung durch analoge oder digitale Auswertung von Impulsschalltestmessungen. Dissertation, Technische Universität Dresden, 1979.

[Sch81] Schroeder, M. R.: Modulation transfer functions: definition and measurement. Acustica, Vol. 49(1981), No. 3, S. 179 – 182.

[SD82] Secrest, B. und Doddington, G.: Postprocessing Techniques for Voice Pitch Trackers. Proc. ICASSP 1982, New York, USA, 1982, S. 172 – 175.

[Sel84] Selkirk, E.: On the major class features and syllable theory. In: Aronoff, M. und Oehrle, R. T. (Eds.): Language Sound Structure. Studies in Phonology presented to Morris Halle by his Teachers and Students. Cambridge, Mass.: MIT Press, 1984, S. 107 – 113.

[SFB01] Stahl, V.; Fischer, A. und Bippus, R.: Acoustic synthesis of training data for speech recognition in living room environments. Proc. ICASSP 2001, Salt Lake City, USA, 2001, S. 285 – 288.

[SGK06] Sehr, A.; Greß, O. und Kellermann, W.: Synthetisches Multicondition-Training zur robusten Erkennung verhallter Sprache. Proc. ITG-Fachtagung Sprachkommunikation, Kiel, 2006.

[SK08] Sehr, A. und Kellermann, W.: Towards Robust Distant-Talking Automatic Speech Recognition in Reverberant Environments. In: Hänsler, E. und Schmidt, G. (Eds.): Speech and Audio Processing in Adverse Environments. Springer-Verlang Berlin Heidelberg, 2008, ISBN 978-3-540-70601-4, S. 679 – 728.

[SI73] Sugiyama, K. und Irii, H.: Comparison of the sound pressure radiation from a prolate spheroid and the human mouth. Acustica, Vol. 73(1991), No. 5, S. 271 – 276.

[SL74] Schmidt, W. und Lehmann, U.: Eignung von Hallabstand oder Hallmaß zur objektiven Bestimmung des Raumeindrucks. Zeitschrift für elektrische Informations- und Energietechnik, Vol. 4(1974), S. 161 – 168.

[Spe84] Specker, P.: A powerful post-processing algorithm for time-domain pitch trackers. Proc. ICASSP 1984, San. Diego, USA, 1984, S. 77 – 80.

[SPW96] Subramaniam, S.; Petropulu, A. und Wendt, C.: Cepstrum-based deconvolution for

speech dereverberation. IEEE Transactions on Speech and Audio Processing, Vol. 4(1996), No. 5, S. 392 – 396.

[SR84] Schmidt, W. und Reichardt, W.: Raumaustik. In: Fasold, W.; Kraak, W.; Schirmer, W. (Hrsg.): Taschenbuch Akustik. Teil 2. VEB Verlag Technik Berlin, 1984, S. 1186 – 1277.

[SS92] Singer, H. und Sagayama, S.: Pitch dependent phone modelling for HMM based speech recognition. Proc. ICASSP 1992, San Francisco, USA, 1992, S. 273 – 276.

[ST99] Schukat-Talamazzini, E. G.: Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen. Vieweg-Verlag, Braunschweig Wiesbaden, 1999, ISBN-13 978-3528054922.

[STFP96] Sachs, J.; Thomä, R.; Friedrich, J.; Peyerl, P.: Vergleichende Untersuchungen zum Einsatz ausgewählter Testsignale in der akustischen Meßtechnik. Sonderdruck. Technische Universität Ilmenau, Fachgebiet Elektronische Meßtechnik und MEODAT GmbH Ilmenau, 1996.

[SU86] Schreiner, C. E. und Urbas, J. V.: Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). Hearing Research, Vol. 21(1986), No. 3, S. 227 – 241.

[SWKN08] Sehr, A.; Wen, J. Y. C.; Kellermann, W. und Naylor, P. A.: A Combined Approach for Estimating a Feature-Domain Reverberation Model in Non-diffuse Environments. Proc. IWAENC 2008, Seattle, USA, 2008.

[SZK06] Sehr, A.; Zeller, M. und Kellermann, W.: Hands-free speech recognition using a reverberation model in the feature domain. Proc. EUSIPCO 2006, Florence, Italien, 2006.

[Tal95] Talkin, D.: A robust algorithm for pitch tracking (RAPT). In: Kleijn, W. B. und Paliwal, K. K. (Eds.): Speech Coding and Synthesis, Elsevier Science, 1995, S. 495 – 518.

[Tay94] Taylor, F. J.: Principles of Signals and Systems. MacGraw-Hill, Inc., New York, 1994.

[Thi53] Thiele, R.: Richtungsverteilung und Zeitfolge der Schallrückwürfe in Räumen. Acustica 3(1953), S. 291 – 302.

[TLK93] Tohyama, M. Lyon, R. H. und Koike, T.: Source waveform recovery in a reverberant space by cepstrum dereverberation. Proc. ICASSP 1993. Minneapolis, USA, 1993, S. 157 – 160.

[TSM79] Takagi, T.; Seiyama, N. und Miyasaka, E.: A method for pitch extraction of speech signal using autocorrelation functions through multiple window length (in Japanisch). IEICE Transactions on Acoustics, Vol. J80-A(1997), No. 9, S. 1341 – 1350.

[TW09] Tschöpe, C und Wolff, M.: Statistical Classifiers for Structural Health Monitoring. IEEE Sensors Journal, Special Issue on Sensor Systems for Structural Health Monitoring, 2009 (im Druck).

[UFSA04] Unoki, M.; Furukawa, M.; Sakata, K. und Akagi, M.: An improved method based on the MTF concept for restoring the power envelope from a reverberant signal. Journal of Acoustic Science & Technology, Vol. 25(2004), No. 4, S. 232 – 242.

[UH08a] Unoki, M. und Hiramatsu, S.: Blind estimation method of reverberation time based on concept of modulation transer function. Proc. Acoustics 2008, Paris, 2008, S. 4489 – 4494.

[UH08b] Unoki, M. und Hiramatsu, S.: MTF-based method of blind estimation of reverberation time in room acoustics. Proc. EUSPICO 2008, Lausanne, Schweiz, 2008.

[UH08c] Unoki, M. und Hosorogiya, T.: Estimation of fundamental frequency of reverbe-

rant speech by utilizing complex cepstrum analysis, Journal Signal Processing, Vol. 12(2008), No. 1, S. 31 – 44.

[UHI08] Unoki, M.; Hosorogiya, T. und Ishimoto, Y.: Comparative evaluation of robust and accurate F0 estimates in reverberant environments. Proc. ICASSP 2008, Las Vegas, 2008, S. 4569 – 4572.

[UPMH08] Unoki, M.; Petrick, R.; Mittal, A. und Hoffmann, R.: Effects of Room Reverberation on Robust and Accurate F0 Estimates. Technical Report of IEICE, Morioka, Iwate, Japan, 2008, S. 1 – 6.

[USA03] Unoki, M.; Sakata, K. und Akagi, M.: A speech dereverberation method based on the MTF concept. Proc. EUROSPEECH 2003, Geneva, Schweiz, 2003, S. 1417 – 1420.

[USFA04] Unoki, M.; Sakata, K.; Furukawa, M. und Akagi, M.: A speech dereverberation method based on the MTF concept in power envelope restoration. Journal of Acoustic Science & Technology, Vol. 25(2004), No. 4, S. 243 – 254.

[UTA05] Unoki, M.; Toi, M. und Akagi, M.: Development of the MTF-based speech dereverberation method using adaptive time-frequency division. Proc. Forum Acusticum 2005, Budapest, Ungarn, 2005, S. 51 – 56.

[VB94] Vorländer, M. und Bietz, H.: Comparison of Methods for Measuring Reverberation Time. Acustica, Vol. 80(1994), S. 205 – 215.

[VHH98] Vary, P.; Heute, U.; Hess, W.: Digitale Sprachsignalverarbeitung. B. G. Teubner Stuttgart, 1998, ISBN-10: 3519061651.

[VM90] Varga, A. P. und Moore, R. K.: Hidden Markov model decomposition of speech and noise. Proc. ICASSP 1990, Albuquerque, USA, 1990, S. 845 – 848.

[Wah00] Wahlster, W. (Ed.): Verbmobil – Foundations of speech-to-speech translations. Springer-Verlag Berlin Heidelberg, 2000, ISBN 3-540-67783-6.

[Wer08] Werner, S.: Sprachsynthese und Spracherkennung mit gemeinsamen Datenbasen, Sprachmodell und Aussprachemodellierung. Dissertation, Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, 2008.

[Wes96] Westendorf, C. M.: Melfilter v1.1. – ein Signalanalyseprogramm. Verbmobil. Technisches Dokument 40, Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, 1996.

[WN06] Wen, J. Y. C. und Naylor, P.: An evaluation measure for reverberant speech using tail decay modelling. Proc. EUSIPCO 2006, Florence, Italien, 2006, S. 1 – 4.

[WPWH07] Wittenberg, S.; Petrick, R.; Wolff, M. und Hoffmann, R.: Einkanalige Störgeräuschunterdrückung zur Steigerung der Worterkennungsrate eines Spracherkenners. In: Studientexte zur Sprachkommunikation, Vol. 33, Tagungsband der 18. Konferenz für Elektronische Sprachsignalverarbeitung (ESSV), TUDpress, Cottbus, 2007, ISBN-13 978-3940046-40-6, S. 52 – 59.

[WS06] Wunsch, G. und Schreiber, H.: Analoge Systeme. TUDpress Lehrbuch, 2006, ISBN 3-938863-67-6.

[Wu03] Wu, M.: Pitch tracking and speech enhancement in noisy and reverberant environments. Dissertation, Ohio State University, 2003.

[WW05] Wu, M. und Wang, D.: A two-stage algorithm for enhancement of reverberant speech. Proc. ICASSP 2005, Philadelphia, USA, 2005, S. 1085 – 1888.

[WW06] Wu, M. und Wang, D.: A two-stage algorithm for enhancement of reverberant speech. IEEE Transactions Speech and Audio Processing, Vol. 14(2006), No. 3, S. 774 – 784.

[Yeg98] Yegnanarayana, B.: Enhancement of reverberant speech using LP residual. Proc.

ICASSP 1998, Seattle, USA, 1998, S. 405 – 408.

[Yeg02] Yegnanarayana, B.: Speech enhancement using excitation source information. Proc. ICASSP 2002, Orlando, USA, 2002, S. 541 – 544.

[YHM07] Yoshioka, T., Hikichi, T. und Miyoshi, M.: Dereverberation by using time-variant nature of speech production system. EURASIP Journal on Advances in Signal Processing, Vol. 2007, Article ID 65698.

[YJM96] Ying, G.; Jamieson, H.; Mitchell, C.: A Probabilistic Approach to AMDF Pitch Detection. Proc. ICSLP 1996, Philadelphia, USA, 1996, S. 1201 – 1204.

[YM00] Yegnanarayana, B. und Murthy, P. S.: Enhancement of reverberant speech using LP residual signal. IEEE Transactions on Speech and Audio Processing, Vol. 8(2000), No. 3, S. 267 – 281.

[ZF67] Zwicker, E. und Feldtkeller, R.: Das Ohr als Nachrichtenempfänger. Hirzel-Verlag, Stuttgart, 1967, ASIN B0000BUCGM.

[ZF90] Zwicker, E.; Fastl, H.: Psychoacoustics – Facts and Models. Springer-Verlag Berlin Heidelberg New York, 1990, ISBN 3-540-52600-5.

Zum Verfasser

Rico Petrick, gelernter Energieelektroniker (Kraftwerk Boxberg, VEAG, heute Vattenfall), studiert von 1996 – 2001 Elektrotechnik an der Hochschule für Technik und Wirtschaft Dresden (FH) sowie 2000 an der University of Technology, Sydney (UTS), Australien. 1998/99 arbeitet er als Praktikant der bei der Firma SysMik GmbH Dresden in der Entwicklung. Von 2001 bis 2004 ist er als Entwicklungsingenieur für Sprachtechnologieprodukte bei der Firma voice INTER connect GmbH tätig. Schwerpunkte seiner Arbeit sind Verfahren der akustischen Signalverarbeitung (Echo- und Geräuschunterdrückung), DSP-Implementierung von Spracherkennern sowie die Gestaltung und Umsetzung von Sprachbenutzerinterfaces für Gerätesteuerungen. Seit 2004 arbeitet er als freiberuflicher Ingenieur für digitale Signalverarbeitung, u. a. für die Firma Hagenuk KMT GmbH in Radeburg. 2006 beginnt er mit der Erarbeitung seiner Dissertation am Institut für Akustik und Sprachkommunikation der Technischen Universität Dresden, wobei Teile der wissenschaftlichen Aktivitäten in jeweils dreimonatigen Forschnungsaufenthalten an der Universitat Poletècnica de Catalunya (UPC), Barcelona, Spanien, dem Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan, sowie an der Russischen Akademie der Wissenschaften, St. Petersburg, Russland, durchgeführt werden. Die vorliegende Dissertation baut thematisch auf den früheren beruflichen Schwerpunkten des Verfassers auf und behandelt ein bislang nahezu unbeleuchtetes Problem für die Robustheit von embedded Spracherkennern: die Störung durch Raumhall.