Artificial Bandwidth Extension of Telephone Speech Signals Using Phonetic *A Priori* Knowledge

Von der Fakultät für Elektrotechnik, Informationstechnik, Physik
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung der Würde
eines Doktor-Ingenieurs (Dr.-Ing.)
genehmigte

Dissertation

von

Patrick Marcel Bauer
aus Stuttgart - Bad Cannstatt

eingereicht am:	1. April 2016
mündliche Prüfung am:	14. Oktober 2016

Referent:	Prof. Dr.-Ing. Tim Fingscheidt Technische Universität Carolo-Wilhelmina zu Braunschweig
Korreferent:	Prof. D.Sc. (Tech.) Paavo Alku Aalto University, Espoo, Finland
Prüfungsvorsitzender:	Prof. Dr.-Ing. Ulrich Reimers Technische Universität Carolo-Wilhelmina zu Braunschweig

2017

Mitteilungen aus dem Institut für Nachrichtentechnik der
Technischen Universität Braunschweig

Band 49

Patrick Marcel Bauer

Artificial Bandwidth Extension of Telephone Speech Signals Using Phonetic *A Priori* Knowledge

Shaker Verlag
Aachen 2017

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Braunschweig, Techn. Univ., Diss., 2016

Editor of this volume:

Prof. Dr.-Ing. Tim Fingscheidt
Institute for Communications Technology
Technische Universität Braunschweig
Schleinitzstrasse 22
38106 Braunschweig
Germany
e-mail: fingscheidt@ifn.ing.tu-bs.de
phone: +49-531-391-2485
fax: +49-531-391-8218

Printed in Germany.

ISBN 978-3-8440-5072-1
ISSN 1865-2484

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Ich möchte diese Dissertation gerne meinem Vater Dr. Eberhard Bauer widmen, der im Jahr 2003 aufgrund einer bösartigen Erkrankung leider viel zu früh von uns gegangen ist.

Danksagung

Diese Dissertation habe ich während meiner wissenschaftlichen Tätigkeit am Institut für Nachrichtentechnik (IfN) der Technischen Universität Braunschweig angefertigt. Ich möchte hiermit allen Professoren, Emeriti, Mitarbeitern aus Technik und Verwaltung, wissenschafftlichen Mitarbeitern sowie studentischen und wisschenschaftlichen Hilfskräften, die mich am IfN begleitet haben, ganz herzlich für die gemeinsame Zeit bedanken. Ebenfalls danke ich allen externen Forschungskollegen und Projektpartnern für die im Rahmen dieser Arbeit erfolgreich durchgeführten wissenschaftlichen Kooperationen und Veröffentlichungen.

Zu tiefstem Dank verpflichtet bin ich dem Referenten der Promotionskommission, meinem Doktorvater Prof. Dr.-Ing. Tim Fingscheidt. Ich möchte mich herzlichst bei Ihnen für die – aus fachlicher wie auch menschlicher Sicht – großartige Zusammenarbeit über all die Jahre bedanken. Sie haben mich stets kompetent betreut und vertrauensvoll unterstützt. Unsere fruchtbaren Diskussionen mit Ihren innovativen Anregungen und wertvollen Ratschlägen waren ein wesentlicher Bestandteil für die Entstehung meiner Disseration. Prof. D.Sc. (Tech.) Paavo Alku möchte ich meinen Dank zum Ausdruck bringen, dass er die Rolle des Korreferenten in der Promotionskommission übernommen und für meine Promotionsprüfung die lange Anreise aus Finnland auf sich genommen hat. Ein herzliches Dankeschön für die Übernahme des Vorsitzes der Promotionskommission gilt Prof. Dr.-Ing. Ulrich Reimers. Ich möchte Ihnen aber auch ganz persönlich in Ihrer Rolle als Leiter des IfN für Ihr bewundernswertes Engagement danken, mit dem Sie allen Mitarbeitern und Studenten am IfN stets auf vorbildlichste Weise begegnen. Insbesondere während meiner Zeit als Oberingenieur des IfN durfte ich dies eindrucksvoll erfahren.

Ohne familiäre Unterstützung wäre meine Dissertation nicht möglich gewesen. Zunächst möchte ich mich dafür bei meiner Mutter, ihrem Lebensgefährten und meinem Bruder bedanken. Bei meiner Frau und meinen drei Kindern bedanke ich mich zudem auch für die liebevolle Rücksichtnahme, die sie mir trotz zahlloser Überstunden entgegen gebracht haben. Dank des medizinischen Fortschritts und unseres starken Familienzusammenhalts konnte sich meine tapfere Tochter von ihrer Leukämieerkrankung erholen. Wir beten zu Gott, dass sie diese Krankheit endgültig besiegt hat und wir von weiteren Rückschlägen verschont bleiben.

Braunschweig im Januar 2017 Patrick Bauer

Kurzfassung

Der ursprünglich durch analoge Übertragungstechniken bedingte schmalbandige Frequenzbereich von Telefonsprachsignalen führt auch in den heutigen digitalen Telefonie-Systemen noch häufig zu akustischen Einschränkungen. Er verursacht dumpf klingende Telefongespräche mit verminderter Sprachverständlichkeit und -qualität. Mittels Verfahren zur künstlichen Sprachbandbreitenerweiterung können fehlende Frequenzkomponenten geschätzt und rekonstruiert werden. Allerdings leidet eine künstlich erweiterte Sprachbandbreite typischerweise unter störenden Artefakten. Besonders anfällig dafür sind die aus dem Schmalbandspektrum schwer zu schätzenden und daher mit anderen Phonemen sowie Sprachpausen leicht zu verwechselnden Frikative /s/ und /z/. Diese Arbeit macht sich phonetisches *A-Priori*-Wissen zu Nutze, um die Leistungsfähigkeit der künstlichen Bandbreitenerweiterung zu optimieren. Sowohl dem vorab offline durchzuführenden Trainingsprozess als auch dem später folgenden eigentlichen Verarbeitungsprozess sollen dadurch wichtige Phoneminformationen zur Verfügung gestellt werden. Da der vorausgehende Trainingsprozess keinerlei Online-Anforderungen stellt, kann phonetisches *A-Priori*-Wissen hierfür verfügbar gemacht werden. Die Verfügbarkeit im späteren Verarbeitungsprozess hängt jedoch von den Online-Anforderungen der jeweiligen Anwendung ab.

In dieser Arbeit werden die beiden Hauptanwendungsfelder der künstlichen Bandbreitenerweiterung behandelt. Auf der einen Seite sollen existierende Telefonsprachdatenbanken in der Bandbreite erweitert werden, um Telefonie-basierte breitbandige Sprachdialogsysteme damit trainieren zu können. Dabei kommt die künstliche Bandbreitenerweiterung vor dem Spracherkennungstraining offline zum Einsatz und benötigt somit für diese Mensch-Maschine-Anwendung (d.h. Telefongespräch mit automatischem Spracherkenner) keine Online-Fähigkeit. Phonetisches *A-Priori*-Wissen kann daher nutzbar gemacht werden. Auf der anderen Seite sollen schmalbandige Telefonsprachdienste künstlich bandbreitenerweitert werden, um deren Verständlichkeit und Qualität zu verbessern. Diese Anwendung von Mensch zu Mensch (d.h. Telefongespräch mit anderem Gesprächspartner) muss online-fähig sein. Daher ist eine geeignete Schätzung des phonetischen *A-Priori*-Wissens erforderlich. Das im Rahmen dieser Arbeit entwickelte Verfahren zur künstlichen Bandbreitenerweiterung konnte seine Leistungsfähigkeit im Vergleich zum Stand der Technik für beide Anwendungsfelder erfolgreich unter Beweis stellen.

Abstract

The narrowband frequency range of telephone speech signals originally caused by former analog transmission techniques still leads to frequent acoustical limitations in today's digital telephony systems. It provokes muffled sounding phone calls with reduced speech intelligibility and quality. By means of artificial speech bandwidth extension approaches, missing frequency components can be estimated and reconstructed. However, the artificially extended speech bandwidth typically suffers from annoying artifacts. Particularly susceptible to this are the fricatives /s/ and /z/. They can hardly be estimated based on the narrowband spectrum and are therefore easily confusable with other phonemes as well as speech pauses. This work takes advantage of phonetic *a priori* knowledge to optimize the performance of artificial bandwidth extension. Both the offline training part conducted in advance and the main processing part performed later on shall be thereby provided with important phoneme information. As the preceding training part does not require online processing, phonetic *a priori* knowledge can be made available. But its availability during the later processing part depends on the online requirements of the particular application.

In this work, the two main application areas of artificial bandwidth extension are addressed. On the one hand, existing telephone speech databases shall be upgraded in bandwidth to be able to train telephony-based wideband interactive voice response systems. For this purpose, the artificial bandwidth extension takes place offline before the speech recognition training and does therefore not require for this human-to-machine application (i.e., telephone conversation with automatic speech recognizer) any online capabilities. Consequently, phonetic *a priori* knowledge can be exploited. On the other hand, narrowband telephone speech services shall be artificially extended in bandwidth to enhance their intelligibility and quality. This human-to-human application (i.e., telephone conversation with another conversational partner) needs to be online-capable. Thus, an appropriate estimation of the phonetic *a priori* knowledge is necessary. The artificial bandwidth extension approach developed within the scope of this work could successfully demonstrate its abilities for both application areas in comparison with the state of the art.

Contents

Danksagung **i**

Kurzfassung **iii**

Abstract **v**

1 Introduction **1**

1.1 Motivation of ABE 1

1.2 ABE History and State of the Art 4

1.3 Contribution of this Work 6

1.4 Outline 8

2 ABE Framework Exploiting Phonetic *A Priori* Knowledge **11**

2.1 ABE Training with Phonetic Support 12

2.1.1 Frame Conversion 13

2.1.2 *Supervised* CB Training 15

2.1.3 LDA Training 19

2.1.4 HMM Training 25

2.2 ABE Processing with Phonetic Support 28

2.2.1 Phonetic Weighting of Observation Likelihoods 30

2.2.2 HMM Decoder 32

2.2.3 Estimation of UB Cepstral Envelope 37

2.2.4 Conversion to UB Spectral Envelope 38

2.2.5 WB Spectral Assembly 39

2.2.6 Conversion to WB LPC Coefficients 40

2.2.7 Interpolation 40

2.2.8 LP Analysis and Synthesis Filtering 41

2.2.9 Residual Signal Extension 43

2.2.10 ABE Postfiltering 46

2.3 Summary 47

3 Human-to-Machine ABE Application: Offline ABE for Training of WB Telephony ASR Systems 49

3.1 Preliminary Phoneme Recognition Experiments . . . 50
3.1.1 Experimental Setup . . . 51
3.1.2 Experimental Results . . . 52
3.2 Phonetically Motivated CB Design for ABE . . . 53
3.2.1 CB with Multiple Phoneme Classes . . . 53
3.2.2 CB with One Phoneme Class . . . 59
3.3 Modification of State Transition Probabilities for ABE . . . 60
3.3.1 Smoothing of State Transitions . . . 61
3.3.2 Attenuation of State Transitions . . . 62
3.3.3 Boosting of State Transitions . . . 62
3.3.4 LSD Performance Evaluation . . . 62
3.4 Large-Vocabulary ASR Experiments with ABE . . . 65
3.4.1 Setup of Large-Vocabulary ASR Experiments . . . 65
3.4.2 ASR Baseline Experiments . . . 69
3.4.3 ABE-Based ASR Experiments . . . 71
3.4.4 Discussion . . . 76
3.5 Summary . . . 77

4 Human-to-Human ABE Application: Online ABE for Enhancement of NB Telephone Speech Services 79

4.1 Preliminary Syllable Articulation Tests with ABE . . . 80
4.1.1 Experimental Setup . . . 80
4.1.2 Experimental Results . . . 84
4.2 ABE Optimization for Speech Quality Enhancement . . . 89
4.2.1 Overview . . . 90
4.2.2 ANN-Based HMM Support . . . 91
4.2.3 SPD-Based Speech Pause Extension . . . 102
4.2.4 Temporal LPC Smoothing . . . 104
4.2.5 Anti-Aliasing of LB Spectrum . . . 105
4.3 Speech Quality Assessment . . . 107
4.3.1 Subjective Speech Quality Tests . . . 109
4.3.2 Instrumental Speech Quality Measurements . . . 116
4.3.3 Discussion . . . 120
4.4 Summary . . . 122

5 Conclusions and Outlook 125

A Phoneme Alphabet SAMPA-D-VMlex **129**

B Statistical Analysis **131**

B.1 Confidence Analysis 131

B.2 Correlation Analysis 132

List of Symbols **135**

List of Abbreviations **139**

Bibliography **143**

Own Publications **163**

Chapter 1

Introduction

Over decades, people have got used to conventional telephony. Did anybody complain about the muffled sound of phone calls? In fact, many hearing-impaired persons severely suffer from it and therefore try to avoid telephone conversations (Kepler et al., 1992). Unfortunately, they cannot compensate their hearing loss by following the lip movement of their conversational partners over the phone. Additionally, background noise makes the communication even more difficult. A volume increase to amplify the speech level directly elevates the noise level as well. In spite of their technological progress, hearing aids and cochlear implants are neither able to offer their users fully relaxed telephone conversations. However, also persons with normal hearing capabilities commonly adapt their way of speaking when having a phone call. Words that can hardly be understood without context information, such as proper names or terms in a foreign language, are often spelled instinctively by means of the telephone alphabet. Hence, there is a great demand for making the muffled sounding conventional telephony brighter, and in consequence also more intelligible.

1.1 Motivation of ABE

Depending on their physiological conditions, lifestyle, and age, humans are capable of perceiving acoustic signals from 0.02 kHz up to 20 kHz (Fastl and Zwicker, 2007, Chap. 2). While music ranges approximately between 0.04 kHz and 10 kHz, speech covers a frequency range of about $0.1 \ldots 7$ kHz (Fastl and Zwicker, 2007, Fig. 2.1). The former analog speech transmission has limited the acoustic bandwidth of conventional telephony up to $0.3 \ldots 3.4$ kHz (Vary and Martin, 2006, Sec. 10.1). Despite the employment of digital speech transmission, a limitation below 4 kHz is often still present nowadays, due to the commonly used sampling rate of $f_s' = 8$ kHz. The question is how much the limited acoustic bandwidth of conventional telephony impacts both speech intelligibility and quality.

Huang et al. (2001, Sec. 9.3.1) empirically predicted for automatic speech recognition (ASR) that a sampling rate reduction from 16 kHz – i.e., wideband (WB) speech – to 8 kHz – i.e., narrowband (NB) speech – would lead to a relative word error rate (WER) increase of about 25 %. This empirical prediction has been confirmed in practice, e.g., by Nadeu and Tolos (2001) as well as Macho and Cheng (2003). Obviously, the recognition performance, which serves as an 'intelligibility' measure for speech recognizers, suffers considerably from the lack of spectral content above 4 kHz arising from the reduced sampling rate. Fundamental experiments on *human* syllable articulation performed by French and Steinberg (1947, Fig. 12) reveal comparable results[1]. The portion of correctly identified contextless syllables decreases significantly from ≈ 98 % to ≈ 90 %, when reducing the cut-off frequency of lowpass-filtered speech from 7 kHz to 3.4 kHz. However, a cut-off frequency increase of highpass-filtered speech from 0.1 kHz to 0.3 kHz turns out to be negligible. It can be therefore concluded that exclusively the missing high frequencies 3.4 ... 7 kHz are responsible for the impaired human speech intelligibility in conventional telephony.

Krebber (1995, Fig. 5.6) carried out speech quality tests with human subjects. The subjective ratings are based on a 5-point listening-quality scale from 1 (bad) to 5 (excellent) and averaged yielding a mean opinion score (MOS) (ITU-T P.800, 1996, Annex B). By decreasing the cut-off frequency of lowpass-filtered speech from 7 kHz to 3.4 kHz, the resulting MOS is reduced by about 0.75 points from ≈ 4.5 to ≈ 3.75. A further MOS reduction by about 0.55 points to ≈ 3.2 arises from bandpass filtering with high and low cut-off frequencies of 3.4 kHz and 0.3 kHz, respectively. Hence, the limited acoustic bandwidth of conventional telephony turns out to degrade the perceived speech quality significantly by about 1.3 MOS points. In contrast to speech intelligibility, not only the high frequencies (> 3.4 kHz) but also the low ones (< 0.3 kHz) have a considerable impact on speech quality.

More and more mobile as well as IP-based telephone networks support high definition (HD) telephony – mostly promoted under the name of *HD Voice*[2] – offering WB speech with an acoustic bandwidth of 0.05 ... 7 kHz (Ferraz de Campos Neto and Järvinen, 2006; Fingscheidt, 2012). WB telephone speech services require a sampling rate of $f_s = 16$ kHz and the use of a WB speech codec[3]. Furthermore, the following requirements need to be fulfilled by the users for successfully establishing an end-to-end WB call: Both conversational partners must employ a WB-capable phone, have the same network provider, and be located in a

[1] For more detail information please also refer to the extensive investigations of Fletcher and Galt (1950).

[2] As of May 2016, 164 mobile network operators commercially launch *HD Voice* in 88 countries (Global Mobile Suppliers Association, 2016).

[3] Today, several WB speech codecs exist, such as the sub-band adaptive differential pulse code modulation (ADPCM) speech codec (ITU-T G.722, 1988), the transform domain speech codec for low frame-loss (ITU-T G.722.1, 1999), the WB adaptive multi-rate (AMR) speech codec (3GPP TS 26.190, 2001), the WB embedded pulse code modulation (PCM) speech codec (ITU-T G.711.1, 2008), and the recently standardized enhanced voice services (EVS) speech codec (3GPP TS 26.441, 2014).

WB-capable network of their provider. Unfortunately, an HD telephony interconnectability between mobile and IP-based telephone networks as well as between different providers of the mobile or IP-based telephone network is still not supported (Global Mobile Suppliers Association, 2016). If an *HD Voice* call falls back into the NB mode (e.g., due to a cell handover or roaming (Möller et al., 2009)), from experience, it will not switch over anymore to the WB mode in practice. Obviously, the resulting number of constraints restricts the actual number of HD telephony call setups to a great extent.

Until the complete market penetration of HD telephony, which may still take a long time for the reasons above, *artificial bandwidth extension (ABE)* techniques provide a fallback solution in order to prevent customer dissatisfaction. Implemented at the receiving side of a terminal[4] (i.e., in the loudspeaker path between speech decoder and digital-to-analog (D/A) converter (Jax, 2002, Fig. 1.4)), the user can directly profit from ABE. It therefore provides an important strategic advantage for terminal manufacturers over other speech enhancement algorithms, such as noise reduction or echo compensation, that mainly represent uplink features serving the conversational partner at the far end. Besides the terminal-specific application, ABE techniques can be also employed inside the infrastructure of heterogeneous telephone networks (Jax, 2002, Sec. 1.2).

From a technical point of view, ABE tries to enhance the speech intelligibility and quality reduced by the acoustic bandwidth limitation of the conventional telephony by recovering the lost speech bandwidth in a 'blind' way, i.e., only based on the NB speech data and *a priori* knowledge. In contrast, other methods like bandwidth extension (BWE) or spectral band replication (SBR) are integrated into a particular speech or audio codec (3GPP TS 26.190, 2001; 3GPP TS 26.404, 2004; 3GPP TS 26.290, 2012; Geiser et al., 2005b, 2007) to exploit side information being additionally transmitted over the transmission channel with a low bit rate (Vary and Martin, 2006, Sec. 10.2). This concept aims at reducing the data rate that would be required for a transmission of full WB speech. Due to the use of additional side information, BWE and SBR are expected to attain a better performance than ABE, but at the expense of a strong dependency on the respective codec. To tackle this problem, digital watermarking approaches have been developed that hide the auxiliary information inside the NB speech signal or bit stream and therefore allow for backward compatibility (Geiser et al., 2005a). However, this concept always requires a modification of the encoder and – for being able to benefit from the hidden information – also of the decoder. The remainder of this work exclusively focuses on completely codec-independent ABE techniques that do not exploit any additionally transmitted or embedded side information.

[4] Amongst others, landline, mobile, digital enhanced cordless telecommunications (DECT) and voice over IP (VoIP) phones as well as other devices like hearing aids, cochlear implants or their remote control units could be equipped with ABE technology (Bauer et al., 2012).

In principle, an ABE could be employed to estimate either the missing low, and/or the missing high frequencies of band-limited speech signals. As aforementioned, the high frequencies have an impact on both speech intelligibility and quality, while the low frequencies only contribute to the speech quality. As the human ear is extremely sensitive to deviations from the original spectral fine structure at low frequencies (Jax, 2002, Sec. 3.5), an accurate reconstruction of the speaker's fundamental frequency is indispensable to prevent annoying artifacts (Chan and Hui, 1997). However, a robust estimation of the fundamental frequency poses a big challenge particularly when dealing with noisy speech (Shahnaz et al., 2007). Hence, many studies report problems regarding the extension to low frequencies (Jax, 2002; Kalgaonkar and Clements, 2009; Thomas et al., 2010; Pulakka et al., 2012d). Furthermore, a low-band extension appears to be less important when considering the fact that the low cut-off frequency of conventional telephony has been further decreased by the digital speech transmission and that the commonly small-dimensioned loudspeakers of mobile phones can hardly reproduce such low frequencies anyway (at least in the speakerphone mode). The remainder of this work exclusively deals with high-band ABE. Thus, the resulting ABE-enhanced speech signals are solely composed of two frequency bands: The *extended* upper-band (UB) spectrum plus the *available* lower-band (LB) spectrum beneath.

1.2 ABE History and State of the Art

In agreement with Jax (2002, Sec. 1.3) as well as Vary and Martin (2006, Sec. 10.3.1), one of the first ABE proposals was made by Schmidt (1933) using nonlinear distortions. A similar concept based on full-wave rectification, bandpass filtering, and high-frequency boosting was applied by the British Broadcasting Corporation (BBC) to improve the sound quality of telephone contributions in broadcast programs (Croll, 1972). However, a rather poor performance of these analog speech processing techniques was reported.

Croll (1972) already claimed that a differentiation between voiced and unvoiced sounds would be necessary for a realistic high-band extension, but he failed in his attempts to develop a reliable voiced/unvoiced classifier. Patrick (1983) was the first one, who exploited a voicing discrimination (Vary and Martin, 2006, Sec. 10.3.1) and thereby made the ABE adaptive (Jax, 2002, Sec. 1.3). In spite of this adaptation, his simple frequency-domain shifting and scaling approach still produced synthetically sounding speech with artifacts.

Dietrich (1984) proposed another rudimentary signal processing technique for ABE. He created high frequency components by means of filter aliasing, which can be easily combined with a sample rate or D/A conversion (Yasukawa, 1995). However, this concept only provides a reasonable performance for an extension of frequencies above 8 kHz (Hänsler and Schmidt, 2008, Sec. 5.4.1). Hence, it has been further developed for audio coding (Vary and Martin,

2006, Sec. 10.3.1), e.g., by Larsen et al. (2002).

The source-filter model of human speech production (Flanagan, 1972) was increasingly taken into account with the beginning of the nineties (Cheng et al., 1992, 1994; Carl, 1994; Carl and Heute, 1994; Yoshida and Abe, 1994; Avendano et al., 1995; Iyengar et al., 1995). This innovation represents a milestone in the history of ABE. It was the starting point of the sophisticated *model-based* ABE algorithms, which basically all rely on two main tasks (Carl, 1994):

1. The creation of a *WB speech excitation signal*, and
2. the creation of a *WB speech spectral envelope*.

In general, these tasks are either processed by means of a *parallel* or *serial* ABE structure (Vary and Martin, 2006, Sec. 10.3.1).

For high-band ABE, the first task is found to be by far less critical than the second one (Carl, 1994; Carl and Heute, 1994; Jax, 2002; Vary and Martin, 2006). There are in principal two reasons that may explain this observation. On the one hand, the human ear is insensitive to deviations from the original spectral fine structure in the UB spectrum (Jax, 2002, Sec. 3.5). On the other hand, the WB speech excitation signal is assumed to be spectrally flat, at least for unvoiced sounds. Nevertheless, different techniques extending the NB speech excitation signal to high frequencies have been reported in literature (Pulakka, 2013, Sec. 4.5.1). Amongst others, they mainly imply *nonlinear distortions* (Jax, 2002, Sec. 3.2), *generation of noise and pitch harmonics* (Jax, 2002, Sec. 3.1), *pitch doubling* (Jax, 2002, Sec. 3.4) as well as *spectral duplication* (Jax, 2002, Sec. 3.3).

The most difficult task for high-band ABE is to accurately estimate the UB speech spectral envelope. Due to insufficient mutual information between the LB and UB spectrum (Nilsson et al., 2002), this estimation represents a one-to-many relationship (Agiomyrgiannakis and Stylianou, 2004). Several techniques have been reported in literature to extend the NB speech spectral envelope (Pulakka, 2013, Sec. 4.5.3). The simplest ones are based on *linear mapping* (Avendano et al., 1995; Nakatoh et al., 1997; Epps and Holmes, 1999; Chennoukh et al., 2001) and *codebook (CB) mapping* (Carl, 1994; Carl and Heute, 1994; Yoshida and Abe, 1994; Chan and Hui, 1997; Enbom and Kleijn, 1999; Epps and Holmes, 1999; Fuemmeler et al., 2001; Iser and Schmidt, 2003; Unno and McCree, 2005; Kornagel, 2006). More sophisticated approaches employ *Gaussian mixture models (GMMs)* (Park and Kim, 2000; Nilsson and Kleijn, 2001; Nilsson et al., 2002; Seltzer et al., 2005; Nour-Eldin and Kabal, 2009; Pulakka et al., 2011; Sunil and Sinha, 2012), *hidden Markov models (HMMs)* (Jax and Vary, 2000; Jax, 2002; Hosoki et al., 2002; Jax and Vary, 2003; Yao and Chan, 2005; Kalgaonkar and Clements, 2008, 2009; Sanna and Murroni, 2009; Thomas et al., 2010; Han et al., 2012; Yagli et al., 2013), or *artificial neural networks (ANNs)* (Tanaka and Hatazoe, 1995; Uncini et al.,

1999; Iser and Schmidt, 2003; Kontio et al., 2007; Pham et al., 2010; Pulakka and Alku, 2011; Pulakka et al., 2014). Furthermore, there are *combined or other methods* (Cheng et al., 1992, 1994; Laaksonen et al., 2005; Yao and Chan, 2006; Ramabadran and Jasiuk, 2008; Laaksonen and Virolainen, 2009; Kalgaonkar, 2011; Katsir et al., 2011, 2012; Pulakka et al., 2014).

Dependencies on the language, dialect, voice or on other pronunciation-specific characteristics of the speaker as well as on the applied codec, acoustical bandwidth and superposed noise of the NB input speech signal still remain a challenge for ABE nowadays (Pulakka, 2013, Sec. 4.7). The state-of-the-art HMM approach of Jax (2002) turned out to be relatively robust against language and speaker variations (Bauer and Fingscheidt, 2008a,b). It forms the initial algorithmic basis of this work.

1.3 Contribution of this Work

In order to develop the baseline ABE algorithm further, some aspects of human speech recognition (HSR) should be considered at first. According to Allen (1994), humans recognize speech by decoding individual speech sounds[5] from the acoustic waveform and interpreting them together via the context. This simplified process involves several cascaded recognition layers (Allen, 1994, Fig. 6), which also serve as a motivation for the main steps of ASR (i.e., feature extraction as well as acoustic and language modeling). The recognition level thereby increases layer by layer from single phonemes over syllables to whole words and sentences. When humans decode a single phoneme, its acoustic waveform is processed and partially recognized in independent spectral articulation bands that get wider with increasing frequency (Allen, 1994, Sec. IV). Each of them makes an additive 5 %-contribution to the so-called articulation index (AI) (French and Steinberg, 1947, Sec. 5). For NB telephone speech, ≈ 15 bands are involved, whereas the remaining ≈ 5 bands belong to the missing high frequencies until 7 kHz (French and Steinberg, 1947, Tab. III). Hence, the potential for ABE in terms of relative AI improvement results in about $\frac{20-15}{15} = \frac{1}{3}$.

However, not all phonemes are equally important for ABE. Most of the phonemes, like vowels, diphthongs, and sonorant consonants, possess considerably more energy in the LB compared to the UB frequency range (Terhardt, 1998, Chap. 7). Please note that in case of

[5]Please note that the term *speech sound* either denotes a *phone* or *phoneme* (Paulus, 1998, Sec. 2.2). Phones are the smallest acoustically distinguishable units of speech independent from a specific language and characterized by a unique pronunciation. In contrast, phonemes are the smallest semantically distinguishable units of speech depending on a particular language and may be pronounced in more than one way (each way of pronunciation represents an allophone). To simplify matters, the terms phone and phoneme are often used synonymously in speech processing literature (the same holds for the corresponding adjectives phonetic and phonemic). This work abstains as well from a strict discrimination and prefers to utilize the term phoneme, so that all transcription symbols are consistently written between slashes /.../.

a negligible UB spectrum, an ABE may entail more risks than chances. Other phonemes, such as plosives and fricatives, reveal a different spectral characteristic. Particularly the unvoiced fricatives /s/, /f/, and /S/ (as well as their voiced counterparts /z/, /v/, /Z/), make a significant UB frequency contribution (Terhardt, 1998, Fig. 7.12). The same holds for the allophones of the purely German 'ch' sound /x/ and /C/, as well as for the unvoiced and voiced fricatives of the purely English 'th' sound /T/ and /D/, respectively. While the noise-like /f/- and /S/-sounds contain noticeable frequency components distributed over the whole WB speech bandwidth, the sharply pronounced /s/-sound only provides a significant spectral content above 3.4 . . . 4 kHz. In general, the ratio between the UB and LB energy is therefore much higher for fricative /s/ (and its voiced counterpart /z/) than for the remaining phonemes. Since the fricatives /s/ and /z/ are hardly distinguishable in the LB spectrum from other phonemes with low energy (e.g., glottal stops or silent/distorted speech pauses), they represent the most critical speech sounds with respect to ABE. This specific problem exemplifies the one-to-many relationship of ABE (Agiomyrgiannakis and Stylianou, 2004) due to the insufficient mutual information between the LB and UB spectrum (Nilsson et al., 2002).

A small phonetic experiment in (Fingscheidt and Bauer, 2013) underlines the importance of the fricatives /s/ and /z/. It involves the phonetically rich sentence "Those answers will be straight forward if you think them through carefully first." originating from the close-talk recordings of the American English speech corpus SpeechDat-Car US (Moreno et al., 2000). The first and third underlined letters stand for a /z/, while the remaining ones represent an /s/. In addition to the WB and NB speech versions, this utterance has been further manipulated by replacing the WB /s/- and /z/-sounds with those of the NB speech signal, and vice versa. The speech quality of the resulting four files was subjectively evaluated by eight subjects in an informal listening test. Interestingly, the NB speech file with implanted WB /s/ and /z/ was rated only 0.12 MOS points worse than the purely WB speech version, and 0.50 MOS points better than the purely NB speech version. Furthermore, it outperformed the WB speech file with implanted NB /s/ and /z/ by 0.69 MOS points. These results suggest the assumption that an appropriate extension of the fricatives /s/ and /z/ is essential for ABE and that most of the other phonemes can more or less get along with the bandwidth limitation. Nevertheless, the extreme case of a digitally driven hard-decision ABE should be prevented, as it may produce audible switching effects.

The fricatives /s/ and /z/ play such an important role for ABE, because their confusion with other phonemes is likely and may involve disturbing artifacts (Bauer et al., 2008). On the one hand, a false rejection of /s/ or /z/ yields an *underestimation*, which typically gives the undesired auditory impression of a 'lisping' speaker (Jax, 2002, Sec. 1.3). On the other hand, a false acceptance of /s/ or /z/ leads to an *overestimation* (Nilsson and Kleijn,

2001), which sounds differently depending on which phoneme is involved. Overestimated voiced sounds commonly provoke 'over-voicing' artifacts. When overestimating unvoiced sounds or speech pauses, 'hissing' artifacts are usually the consequence. It can be assumed that these annoying under- and overestimation artifacts form the biggest obstacle for the commercialization of ABE.

This work aims at reducing the under- and overestimation artifacts. Its main contribution is the development of an ABE framework derived from (Jax, 2002), but exploiting phonetic *a priori* knowledge. For this purpose, the ABE training and processing parts will be both modified to make use of the phonetic support (Bauer and Fingscheidt, 2009a). These modifications take into account that the required phonetic *a priori* information can be made available offline to ABE training (e.g., manually by humans or automatically via forced Viterbi alignment), while the availability for ABE processing depends on the real-time and latency requirements of the given application. This work considers offline as well as online ABE applications (Bauer et al., 2014b,a), whereby the term 'online' is understood to mean real-time capability with certain latency constraints. In the more challenging second case, the ABE processing needs to be online-capable and therefore either does without phonetic support or performs a phonetic estimation in real time. By reducing the under- and overestimation artifacts, both the speech intelligibility and quality shall be improved. Preliminarily, the speech intelligibility will be evaluated by means of automatic phoneme recognition experiments (Bauer et al., 2010d,c) and human syllable articulation tests (Bauer et al., 2010b, 2012, 2013). As it has been still unclear so far, how to reliably assess ABE in terms of speech quality, several instrumental measures and subjective listening tests will be investigated to finally propose a reliable speech quality assessment methodology (Bauer et al., 2014c).

Please note that phonetically motivated approaches have already been used for speech enhancement (Hansen and Pellom, 1997; Das and Hansen, 2012). In the context of ABE, Laaksonen et al. (2005) classified speech frames into phonetic categories based on feature comparisons and thereby selected the cubic spline parameters for spectral shaping. In (Katsir et al., 2011), a phoneme-specific HMM is trained following (Bauer and Fingscheidt, 2009a) and employed for frame classification to select the corresponding WB vocal tract area CB for mapping, as inspired by (Epps and Holmes, 1999).

1.4 Outline

Until now, the main topic of this work has been introduced. The remainder of this work is organized as follows.

In *Chap. 2*, the algorithmic fundamentals on ABE exploiting phonetic *a priori* knowledge

are established. The formulated ABE framework is based on the state-of-the-art HMM approach of Jax (2002) and developed further by several innovations. The most important one is the exploitation of phonetic *a priori* information in support of the ABE training and processing. While Chap. 2 assumes this phonetic support to be available, the subsequently following two chapters demonstrate, how the proposed ABE framework can be used for applications without and with online requirements in practice.

Probably the most intuitive offline ABE application is the spectral restoration of NB speech material originating, e.g., from black box recordings, telephone broadcast contributions, or historical audio archives (Bauer and Fingscheidt, 2009b). Apart from that, *Chap. 3* focuses on a human-to-machine ABE application, which can make use of the numerously available NB telephone speech corpora. The idea behind this is to extend their acoustic bandwidth offline via ABE for the ASR training of interactive voice response (IVR) systems supporting HD telephony services. Due to the lack of WB telephone speech corpora, such a database upgrade is valuable as it prevents expensive and time-consuming speech recordings. Other approaches also utilize existing NB telephone speech data to compensate for the bandwidth mismatch in ASR training (Liao et al., 2003; Seltzer and Acero, 2005, 2007; Karafiat et al., 2007). However, they all require modifications of the employed ASR system affecting the feature extraction and/or acoustic model training. In contrast, the ABE framework proposed in this work operates completely independent from the speech recognizer. As neither the training nor the processing of the offline ABE demands any online requirements, the phonetic *a priori* knowledge can be made available for both.

A human-to-human ABE application, for which the phonetic *a priori* information can be provided offline only to ABE training, is presented in *Chap. 4*. It deals with an online ABE for the enhancement of conventional NB telephone speech services. The creation of a real-time phonetic support for online ABE processing represents the most challenging innovation in this chapter. Further algorithmic modifications are introduced to optimize the proposed ABE in terms of speech quality from a human point of view. This optimization is important, as the previous chapter placed emphasis on improving 'intelligibility' for the purpose of automatic speech recognizers. At the end of this chapter, a reliable speech quality assessment methodology is developed based on the analysis of widely-used subjective listening tests and instrumental measurements.

Finally, *Chap. 5* draws conclusions from the preceding chapters and risks a look into the future work of ABE.

Chapter 2

ABE Framework Exploiting Phonetic *A Priori* Knowledge

According to the last chapter, HSR relies on the detection of single phonemes. A limitation of the acoustic bandwidth arising from the conventional telephony causes a significant speech intelligibility and quality reduction. Phonemes with main spectral components above 4 kHz, such as the unvoiced fricatives /s/, /f/, /S/, and their voiced counterparts /z/, /v/, /Z/, particularly suffer from NB telephone speech. Due to their small energy content below 4 kHz, /s/ and /z/ are considered as the most critical phonemes in terms of ABE. Based on the NB spectrum, they can be easily confused with other phonemes. Thus, it is challenging for an ABE to recognize them correctly. A misrecognition may provoke annoying under- as well as overestimation artifacts. In order to improve speech intelligibility and quality, these artifacts must be prevented. Hence, this chapter formulates an ABE framework after (Bauer and Fingscheidt, 2009a), which is supported by phonetic *a priori* knowledge to allow for a better recognition of critical phonemes. This phonetic support is realized by informing the frame-wise ABE training and processing frame by frame about the current phoneme class. The respective phoneme classes need to be identified by means of frame-aligned phonetic transcriptions that are assumed to be available in advance. A phoneme class may represent just single phonemes or clusters of multiple phonemes, depending on its predefined specification. For the purpose of generalization, this chapter allows for an arbitrary phoneme class specification. Sec. 2.1 and 2.2 explain in detail the statistical framework of phonetically supported ABE training and processing, respectively.

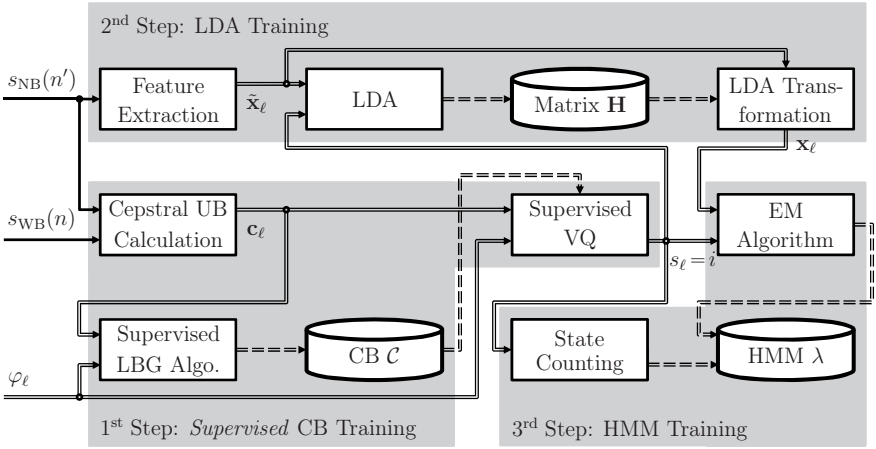


Figure 2.1: Block diagram of ABE training exploiting phonetic *a priori* knowledge.

2.1 ABE Training with Phonetic Support

The employed ABE training is basically conducted in three successive steps according to (Jax, 2002, Sec. 6.3.4). They comprise the trainings of a CB $\mathcal{C}$, a linear discriminant analysis (LDA) matrix $\mathbf{H}$, and a hidden Markov model (HMM) λ. To tackle the problem of the aforementioned ABE artifacts, the original ABE training scheme is developed further by introducing phonetic *a priori* knowledge (Bauer and Fingscheidt, 2009a). Thus, relevant speech sounds can be specifically taken into account.

Fig. 2.1 depicts the block diagram of the proposed ABE training. For each frame ℓ, phonetic information in terms of the phoneme class label φ_ℓ is explicitly exploited during the first training step yielding a *supervised* CB training. Since the derived CB is essential for the subsequent linear discriminant analysis (LDA) and HMM training, the phonetic support implicitly influences also these remaining training steps. In addition to φ_ℓ, temporally aligned WB and NB speech data $s_{\mathrm{WB}}(n)$ and $s_{\mathrm{NB}}(n')$, respectively, serve as input to the ABE training. The sample index n thereby refers to a sampling rate of $f_{\mathrm{s}} = 16$ kHz and n' to $f'_{\mathrm{s}} = 8$ kHz. Both are related to the frame index via $\ell = \lfloor n/N_{\mathrm{s}} \rfloor = \lfloor n'/N'_{\mathrm{s}} \rfloor$, with N_{s} and $N'_{\mathrm{s}} = N_{\mathrm{s}}/2$ denoting the corresponding frame shifts in samples.

Please note that signaling paths represented by single lines in Fig. 2.1 denote a sample-wise processing, whereas a frame-wise processing is characterized by double lines[1]. Obviously, all signals are processed frame by frame. The input speech samples $s_{\mathrm{WB}}(n)$ and $s_{\mathrm{NB}}(n')$ therefore need to be directly converted into frames within the cepstral UB calculation and feature

[1]This convention consistently applies to the remaining ABE block diagrams of this work.

extraction blocks. This frame conversion is initially described based on (P. Bauer, 2007, Sec. 5.1.2), before going into details of the particular ABE training steps.

2.1.1 Frame Conversion

The division of a speech signal into blocks of samples (i.e., speech frames) is typically done by means of a window function of length N_w with the following property:

$$w(n) = 0, \quad \forall\, n < 0,\ n \geqslant N_\mathrm{w}. \tag{2.1}$$

To include context of adjacent frames, a window overlap is used. It provides a look-back and look-ahead by using N_- samples of the preceding frame and N_+ samples of the subsequent frame, respectively. Along with the actual frame length N, the window length results in $N_\mathrm{w} = N_- + N + N_+$.

In case of WB speech data, a conversion into L frames is obtained by shifting $s_\mathrm{WB}(n)$ for each frame ℓ with a frame shift of $N_\mathrm{s} = N$ samples further and applying a multiplication with the window function $w(n)$:

$$s_{\mathrm{WB},\ell}(n) = s_\mathrm{WB}(n + \ell N_\mathrm{s} - N_-) \cdot w(n), \quad n = 0, 1, \ldots, N_\mathrm{w} - 1, \ \ \ell = 0, 1, \ldots, L - 1. \tag{2.2}$$

For initialization purposes, the look-back samples of the zeroth frame may be set to 0. As windowing is expressed by a multiplication in the time domain, it represents a convolution in the frequency domain.

Ideally, the Fourier transform of the window function is a Dirac impulse. However, this would require an infinitely long window in the time domain ($N_\mathrm{w} \to \infty$), which is not feasible for a frame-wise processing. A shortening of the window length intuitively leads to a rectangular, so-called boxcar window

$$w(n) = \begin{cases} 1, & \text{if } n = 0, 1, \ldots, N_\mathrm{w} - 1, \\ 0, & \text{else.} \end{cases} \tag{2.3}$$

Since its amplitude response over the normalized angular frequency Ω

$$\left|W\left(e^{j\Omega}\right)\right| = \frac{|\sin(\Omega N_\mathrm{w}/2)|}{|\sin(\Omega/2)|}, \quad -\pi \leqslant \Omega \leqslant \pi, \tag{2.4}$$

reveals relatively high side lobe levels independent from N_w, undesired spectral leakage effects are the consequence (Proakis and Manolakis, 2007, Sec. 10.2.2). Hence, the choice of the window function is very important.

In contrast to the boxcar window, more sophisticated window functions, such as the Hamming, Hann, or Blackman window, do not contain abrupt discontinuities in the time

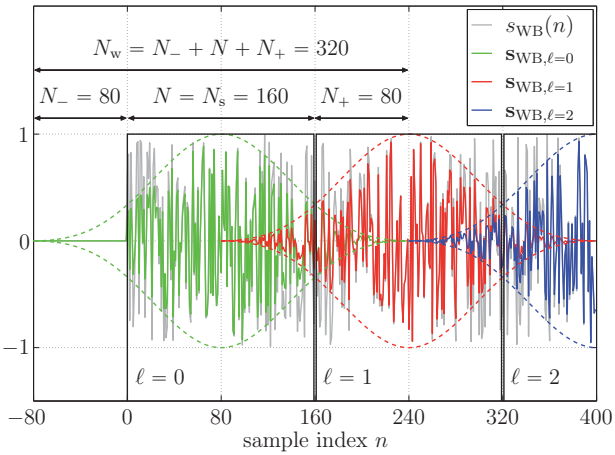


Figure 2.2: Frame conversion of the normalized WB speech data $s_{\mathrm{WB}}(n)$.

domain and therefore provide a higher side lobe attenuation in the frequency domain. The highest side lobe attenuation among typically employed window functions is achieved by the Blackman window (Proakis and Manolakis, 2007, Sec. 10.2.2)

$$w(n) = \begin{cases} 0.42 - 0.5 \cdot \cos\left(\frac{2\pi n}{N_w - 1}\right) + 0.08 \cdot \cos\left(\frac{4\pi n}{N_w - 1}\right), & \text{if } n = 0, 1, \ldots, N_w - 1, \\ 0, & \text{else.} \end{cases} \tag{2.5}$$

In return, it provokes more smoothing in the frequency domain due to its wider main lobe. Anyway, the width of the main lobe can be reduced by increasing N_w.

When making use of a sophisticated window function, a window overlap is recommended to allow for an equal influence of all samples over the frames. In case of no window overlap, samples located near the frame edges would be neglected due to the small weights of the window at these positions.

Fig. 2.2 visualizes the frame conversion of the normalized WB speech data $s_{\mathrm{WB}}(n)$ for the first frames $\ell = 0, 1, 2$. It involves a Blackman window (2.5) with a symmetrical 50 % window overlap using a look-back of $N_- = 80$ samples, a frame length of $N = 160$ samples, and a look-ahead of $N_+ = 80$ samples. This results in a window length of $N_w = 320$ samples. The frame shift is set to $N_s = 160$ samples equal to the frame length. Given the sampling rate of $f_s = 16$ kHz, this corresponds to a frame duration of 10 ms, which is commonly used in speech processing and recognition applications.

The frame conversion of the NB speech data $s_{\mathrm{NB}}(n')$ is done in accordance with (2.2):

$$s_{\mathrm{NB},\ell}(n') = s_{\mathrm{NB}}(n' + \ell N'_s - N'_-) \cdot w(n'), \quad n' = 0, 1, \ldots, N'_w - 1, \quad \ell = 0, 1, \ldots, L - 1. \tag{2.6}$$

It employs a downsampled version of the Blackman window (2.5)

$$w(n') = \begin{cases} 0.42 - 0.5 \cdot \cos\left(\frac{2\pi n'}{N'_\mathrm{w}-1}\right) + 0.08 \cdot \cos\left(\frac{4\pi n'}{N'_\mathrm{w}-1}\right), & \text{if } n' = 0, 1, \ldots, N'_\mathrm{w} - 1, \\ 0, & \text{else.} \end{cases} \tag{2.7}$$

The required parameters are just divided by two, i.e., $N'_- = N_-/2 = 40$, $N'_\mathrm{s} = N_\mathrm{s}/2 = N' = N/2 = 80$, $N'_+ = N_+/2 = 40$, and $N'_\mathrm{w} = N'_- + N' + N'_+ = N_\mathrm{w}/2 = 160$. This allows for a temporal alignment of the NB and WB speech frames / windows.

Please note that the above specified frame conversion is consistently used throughout this work. In contrast, Jax (2002) obviously utilizes non-overlapping frames of 20 ms duration, as stated in (Jax, 2002, p. 14 and p. 20). The explicit use of a window function is mentioned only in (Jax, 2002, Sec. 4.1.2). Apart from that, it can be assumed that a boxcar window is applied in (Jax, 2002).

2.1.2 *Supervised* CB Training (1st Step)

In the 1st step of ABE training depicted in Fig. 2.1, frame-wise UB cepstral envelope vectors $\mathbf{c}_\ell$ are calculated and fed into the well-known Linde-Buzo-Gray (LBG) algorithm for vector quantization (VQ) (Linde et al., 1980). Thus, a CB $\mathcal{C}$ representing various speech sounds in the upper frequency band is trained and utilized to quantize $\mathbf{c}_\ell$. The resulting quantization index i defines the state s_ℓ of the HMM for the remaining training steps. As opposed to (Jax, 2002, Sec. 6.2), both the LBG algorithm and the VQ are supervised by phonetic *a priori* knowledge in terms of the frame-aligned phoneme class labels φ_ℓ (Bauer and Fingscheidt, 2009a). Hence, the obtained CB $\mathcal{C}$ and HMM states $s_\ell = i$ are influenced by this phonetic support. Subsequently, the cepstral UB calculation is described, followed by the supervised LBG algorithm and VQ inspired by (Yu et al., 1990).

Calculation of UB Cepstral Envelope

The required UB cepstral envelope vectors $\mathbf{c}_\ell$ are computed similarly to (Jax, 2002, Sec. 4.1.2) via selective linear prediction (SLP) (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978), as depicted in Fig. 2.3. Frequency bands can be thereby specified flexibly in the frequency domain, which is not the case given a time-domain auto-correlation function (ACF) (Jax, 2002, Sec. 4.1.1). In this work, the frequency band specification is done by means of a cut-off frequency $f_\mathrm{c} \in (0, f'_\mathrm{s}/2]$, which separates the lower and upper frequency bands from each other. This results in an LB and UB frequency range of $[0, f_\mathrm{c}]$ and $(f_\mathrm{c}, f_\mathrm{s}/2]$, respectively. In the special case of $f_\mathrm{c} = f'_\mathrm{s}/2 = 4$ kHz, which represents the default setting in this work, the LB frequencies range from 0 kHz to 4 kHz. This represents the maxi-

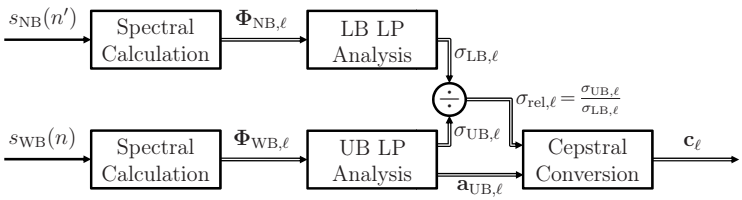


Figure 2.3: Detailed block diagram of UB cepstral envelope calculation as part of the 1st step in Fig. 2.1.

mum bandwidth that can be theoretically provided by NB speech according to the sampling theorem (Rabiner and Schafer, 1978, Sec. 2.4.1).

The spectral calculations based on the WB and NB input speech samples in Fig. 2.3 imply at first a frame conversion according to (2.2) and (2.6), respectively. Afterwards, a short-term discrete Fourier transform (DFT)[2] is applied (Oppenheim and Schafer, 1989, Sec. 8.1)

$$S_{\mathrm{WB},\ell}(k) = \sum_{n=0}^{N_{\mathrm{w}}-1} s_{\mathrm{WB},\ell}(n) \cdot e^{-j2\pi \frac{nk}{N_{\mathrm{w}}}}, \quad k = 0, 1, \ldots, N_{\mathrm{w}} - 1, \tag{2.8}$$

$$S_{\mathrm{NB},\ell}(k') = \sum_{n'=0}^{N'_{\mathrm{w}}-1} s_{\mathrm{NB},\ell}(n') \cdot e^{-j2\pi \frac{n'k'}{N'_{\mathrm{w}}}}, \quad k' = 0, 1, \ldots, N'_{\mathrm{w}} - 1, \tag{2.9}$$

yielding the WB and NB short-term power spectra

$$\Phi_{\mathrm{WB},\ell}(k) = |S_{\mathrm{WB},\ell}(k)|^2, \tag{2.10}$$

$$\Phi_{\mathrm{NB},\ell}(k') = |S_{\mathrm{NB},\ell}(k')|^2, \tag{2.11}$$

with the discrete frequency bin indices k and k' covering the respective frequency band.

Fig. 2.3 subsequently involves a sub-band linear prediction (LP) analysis in order to compute the UB LP filter coefficients $\mathbf{a}_{\mathrm{UB},\ell}$ as well as the corresponding prediction gain factor $\sigma_{\mathrm{UB},\ell}$ (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978). Additionally, the LB prediction gain factor $\sigma_{\mathrm{LB},\ell}$ is computed for normalization purposes. The truncated UB and LB ACFs required for sub-band LP analysis are derived via the short-term inverse partial DFTs (Oppenheim and Schafer, 1989, Sec. 8.1)

$$\phi_{\mathrm{UB},\ell}(\tilde{n}) = \frac{1}{K_{\mathrm{UB}}} \sum_{\tilde{k} \in \tilde{\mathcal{K}}_{\mathrm{UB}}} \Phi_{\mathrm{WB},\ell}\left(\mathrm{k}_{\mathrm{UB}}(\tilde{k})\right) \cdot e^{j2\pi \frac{\tilde{n}\tilde{k}}{K_{\mathrm{UB}}}}, \quad \tilde{n} = 0, 1, \ldots, N_{\mathrm{LP}^{(\mathrm{UB})}}, \tag{2.12}$$

$$\phi_{\mathrm{LB},\ell}(\tilde{n}') = \frac{1}{K_{\mathrm{LB}}} \sum_{\tilde{k}' \in \tilde{\mathcal{K}}_{\mathrm{LB}}} \Phi_{\mathrm{NB},\ell}\left(\mathrm{k}_{\mathrm{LB}}(\tilde{k}')\right) \cdot e^{j2\pi \frac{\tilde{n}'\tilde{k}'}{K_{\mathrm{LB}}}}, \quad \tilde{n}' = 0, 1, \ldots, N_{\mathrm{LP}^{(\mathrm{LB})}}, \tag{2.13}$$

[2] In practice, the DFT can be efficiently computed via a fast Fourier transform (FFT) implementation (Oppenheim and Schafer, 1989, Sec. 9.1).

with the sample indices $\tilde{n}$ and $\tilde{n}'$, respectively. The corresponding sub-band LP orders are denoted by $N_{\mathrm{LP}^{(\mathrm{UB})}} \leqslant K_{\mathrm{UB}}-1$ and $N_{\mathrm{LP}^{(\mathrm{LB})}} \leqslant K_{\mathrm{LB}}-1$. In this work, $N_{\mathrm{LP}^{(\mathrm{UB})}} = 8$ and $N_{\mathrm{LP}^{(\mathrm{LB})}} = 10$ are used. By means of a critical downsampling (Jax, 2002, Sec. 4.1.1), the frequency bin indices $\tilde{k} = 0, 1, \ldots, K_{\mathrm{UB}}-1$ and $\tilde{k}' = 0, 1, \ldots, K_{\mathrm{LB}}-1$, which address only the spectral components of the respective sub-band short-term power spectrum, are mapped to the whole frequency range of $0 \ldots 2\pi$. The numbers of frequency bins are defined by $K_{\mathrm{UB}} = N_{\mathrm{w}} - K_{\mathrm{LB}}$ and $K_{\mathrm{LB}} = 2\lfloor N'_{\mathrm{w}} \cdot f_{\mathrm{c}}/f'_{\mathrm{s}} \rceil$. For convenience, both K_{UB} and K_{LB} are even numbers ensured by the nearest-integer rounding operation $\lfloor \cdot \rceil$ combined with a multiplication by 2. In the special case of $f_{\mathrm{c}} = f'_{\mathrm{s}}/2 = 4$ kHz, they result in $K_{\mathrm{UB}} = K_{\mathrm{LB}} = N'_{\mathrm{w}} = 160$. To critically downsample the sub-band short-term power spectra, two mapping functions are used. On the one hand, the mapping function $\mathrm{k}_{\mathrm{UB}}(\cdot)$ is applied to extract the UB power spectrum from the WB power spectrum $\Phi_{\mathrm{WB},\ell}(k)$ by mapping the domain $\tilde{\mathcal{K}}_{\mathrm{UB}} = \{0, 1, \ldots, K_{\mathrm{UB}} - 1\}$ of frequency bins $\tilde{k}$ to the range $\mathcal{K}_{\mathrm{UB}} = \left\{\frac{K_{\mathrm{LB}}}{2}, \frac{K_{\mathrm{LB}}}{2} + 1, \ldots, N_{\mathrm{w}} - \frac{K_{\mathrm{LB}}}{2} - 1\right\}$ of frequency bins $k = \mathrm{k}_{\mathrm{UB}}(\tilde{k})$:

$$\mathrm{k}_{\mathrm{UB}} \;:\; \tilde{\mathcal{K}}_{\mathrm{UB}} \mapsto \mathcal{K}_{\mathrm{UB}} \;:\; \tilde{k} \mapsto \mathrm{k}_{\mathrm{UB}}(\tilde{k}) = k. \tag{2.14}$$

On the other hand, the LB power spectrum is taken from the NB power spectrum $\Phi_{\mathrm{NB},\ell}(k')$ by means of the mapping function $\mathrm{k}_{\mathrm{LB}}(\cdot)$, which maps the domain $\tilde{\mathcal{K}}_{\mathrm{LB}} = \{0, 1, \ldots, K_{\mathrm{LB}} - 1\}$ of frequency bins $\tilde{k}'$ to the range $\mathcal{K}_{\mathrm{LB}} = \left\{0, 1, \ldots, \frac{K_{\mathrm{LB}}}{2} - 1, N'_{\mathrm{w}} - \frac{K_{\mathrm{LB}}}{2}, \ldots, N'_{\mathrm{w}} - 1\right\}$ of frequency bins $k' = \mathrm{k}_{\mathrm{LB}}(\tilde{k}')$:

$$\mathrm{k}_{\mathrm{LB}} \;:\; \tilde{\mathcal{K}}_{\mathrm{LB}} \mapsto \mathcal{K}_{\mathrm{LB}} \;:\; \tilde{k}' \mapsto \mathrm{k}_{\mathrm{LB}}(\tilde{k}') = k'. \tag{2.15}$$

Based on the resulting truncated ACFs $\boldsymbol{\phi}_{\mathrm{UB},\ell}$ and $\boldsymbol{\phi}_{\mathrm{LB},\ell}$, the required sub-band linear predictive coding (LPC) parameters $\mathbf{a}_{\mathrm{UB},\ell}$, $\sigma_{\mathrm{UB},\ell}$, and $\sigma_{\mathrm{LB},\ell}$ are derived as in a conventional LP analysis, e.g., by efficiently solving the auto-correlation method via the well-known Levinson-Durbin recursion (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978).

The last step of Fig. 2.3 comprises the recursive conversion of the obtained UB LP filter coefficients $\mathbf{a}_{\mathrm{UB},\ell} = [a_{\mathrm{UB},\ell}(1), a_{\mathrm{UB},\ell}(2), \ldots, a_{\mathrm{UB},\ell}(N_{\mathrm{LP}^{(\mathrm{UB})}})]^{\mathrm{T}}$ into linear predictive cepstral coefficients (LPCCs) $\mathbf{c}_\ell = [c_\ell(0), c_\ell(1), \ldots, c_\ell(N_{\mathrm{LP}^{(\mathrm{UB})}})]^{\mathrm{T}}$ (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978):

$$c_\ell(n) = a_{\mathrm{UB},\ell}(n) + \sum_{\nu=1}^{n-1} \frac{\nu}{n} \cdot a_{\mathrm{UB},\ell}(n-\nu) c_\ell(\nu), \quad n = 1, 2, \ldots, N_{\mathrm{LP}^{(\mathrm{UB})}}. \tag{2.16}$$

Thus, the UB spectral envelope represented by $\mathbf{a}_{\mathrm{UB},\ell}$ is transferred into the cepstral domain. For convenience, $\mathbf{c}_\ell$ is therefore referred to as *UB cepstral envelope*. Taking into account (Jax, 2002, Eqs. (4.9) and (4.14)), the zeroth LPCC is defined as

$$c_\ell(0) = \frac{\ln(\sigma^2_{\mathrm{rel},\ell})}{\sqrt{2}}, \tag{2.17}$$

with $\sigma_{\mathrm{rel},\ell} = \sigma_{\mathrm{UB},\ell}/\sigma_{\mathrm{LB},\ell}$ denoting the ratio between the UB and LB prediction gains. This provokes a sub-band energy normalization adapting the level of the UB cepstral envelope.

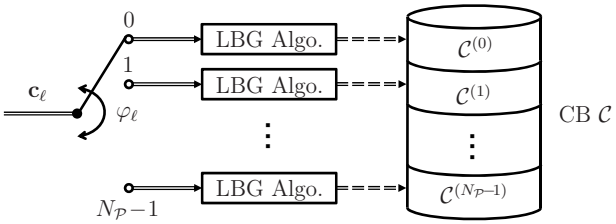


Figure 2.4: Detailed block diagram of the supervised LBG algorithm as part of the 1st step in Fig. 2.1.

Such a level adaptation is indispensable to allow for a robust ABE algorithm. Later on in Sec. 2.2.5, the sub-band energy normalization will be compensated for by the LB prediction gain of the NB speech frames during ABE processing. As in the LB frequency range a mismatch between the spectral characteristics of $s_{\mathrm{NB}}(n')$ and $s_{\mathrm{WB}}(n)$ can be highly expected in practice, e.g., due to channel transmission effects, speech coding, etc., Fig. 2.3 makes use of both speech signals. In contrast, the SLP technique in (Jax, 2002, Fig. 4.3) employs only $s_{\mathrm{WB}}(n)$.

Supervised LBG Algorithm and VQ

Based on the calculated UB cepstral envelope vectors $\mathbf{c}_\ell \; \forall \; \ell = 0, 1, \ldots, L-1$, the required CB for VQ is trained via the well-known LBG algorithm (Linde et al., 1980): After each binary split, the k-means algorithm is used to iteratively improve the CB entries. Due to the binary splitting, the CB size always results in a power of two[3].

As opposed to (Jax, 2002, Sec. 6.2), the CB training via the LBG algorithm is supervised by frame-wise phoneme class labels $\varphi_\ell \in \mathcal{P} = \{0, 1, \ldots, N_\mathcal{P} - 1\}$, with the number of phoneme classes $N_\mathcal{P} \geqslant 1$ denoting the size of the phoneme class alphabet $\mathcal{P}$. In principal, each phoneme class can be devoted to a single phoneme (one-to-one mapping) or a cluster of multiple phonemes (one-to-many mapping). In the special case, where only one phoneme class represents all phonemes, the CB training is identical to (Jax, 2002, Sec. 6.2). While the specific phoneme class assignment is presented in Sec. 3.2 involving various phonetically motivated CB designs, this section focuses on the generalized concept of supervised CB training inspired by (Yu et al., 1990).

Fig. 2.4 depicts the block diagram of the supervised LBG algorithm. For each phoneme class $\varphi \in \mathcal{P}$, a conventional LBG algorithm is individually conducted yielding a sub-CB

[3]Please note that the k-means algorithm can basically be used also without the binary splitting steps of the LBG algorithm. Thus, arbitrary CB sizes are feasible. However, the performance of the k-means algorithm highly depends on the initial CB guess (Linde et al., 1980).

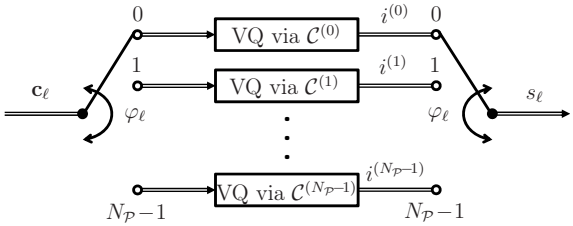


Figure 2.5: Detailed block diagram of the supervised VQ with index output; this is the last operation of the 1st step in Fig. 2.1.

$\mathcal{C}^{(\varphi)} = \left\{\mathbf{c}^{(i^{(\varphi)})} \mid i^{(\varphi)} \in \mathcal{S}^{(\varphi)}\right\}$ of size $N_\mathcal{S}^{(\varphi)}$. Each of its entries $\mathbf{c}^{(i^{(\varphi)})} = E\left\{\mathbf{c}_\ell \mid s_\ell = i^{(\varphi)}\right\}$ thus represents a statistical expectation $E\{\cdot\}$ of all vectors $\mathbf{c}_\ell$ assigned to sub-CB class $i^{(\varphi)}$ (Jax, 2002, Eq. (6.9)). The set of HMM states $\mathcal{S}^{(\varphi)} = \left\{\bar{N_\mathcal{S}}^{(\varphi)}, \bar{N_\mathcal{S}}^{(\varphi)} + 1, \ldots, \bar{N_\mathcal{S}}^{(\varphi)} + N_\mathcal{S}^{(\varphi)} - 1\right\}$ is defined by the indices $i^{(\varphi)}$ of the sub-CB entries. A variable offset $\bar{N_\mathcal{S}}^{(\varphi)} = \sum_{p=0}^{\varphi-1} N_\mathcal{S}^{(p)}$ thereby allows for a correct indexing of the sub-CBs. The allocation of $\mathbf{c}_\ell$ to the phoneme-class-specific data sets is driven by φ_ℓ. A fusion of all sub-CBs yields the final CB $\mathcal{C} = \left\{\mathcal{C}^{(\varphi)} \mid \varphi \in \mathcal{P}\right\} = \left\{\mathbf{c}^{(i)} \mid i \in \mathcal{S}\right\}$ of size $N_\mathcal{S} = \sum_{\varphi\in\mathcal{P}} N_\mathcal{S}^{(\varphi)}$. The indices i of the CB entries $\mathbf{c}^{(i)} = E\{\mathbf{c}_\ell \mid s_\ell = i\}$ thereby define the complete set of HMM states $\mathcal{S} = \left\{\mathcal{S}^{(\varphi)} \mid \varphi \in \mathcal{P}\right\} = \{0, 1, \ldots, N_\mathcal{S} - 1\}$.

By means of the trained CB $\mathcal{C}$, the UB cepstral envelope vector $\mathbf{c}_\ell$ is quantized frame by frame according to Fig. 2.5. Following the supervised LBG algorithm, the VQ makes also use of the phonetic *a priori* knowledge, in contrast to (Jax, 2002, Sec. 6.2). Based on phoneme class label $\varphi_\ell \in \mathcal{P}$, a VQ $\mathbf{Q}^{(\varphi_\ell)}$ is conducted via sub-CB $\mathcal{C}^{(\varphi_\ell)}$ by assigning $\mathbf{c}_\ell$ to the sub-CB entry $\mathbf{c}^{(i^{(\varphi_\ell)})} \in \mathcal{C}^{(\varphi_\ell)}$:

$$\mathbf{c}^{(i^{(\varphi_\ell)})} = \mathbf{Q}^{(\varphi_\ell)}[\mathbf{c}_\ell]\,. \tag{2.18}$$

The resulting quantization index $i^{(\varphi_\ell)} \in \mathcal{S}^{(\varphi_\ell)} = \left\{\bar{N_\mathcal{S}}^{(\varphi_\ell)}, \bar{N_\mathcal{S}}^{(\varphi_\ell)} + 1, \ldots, \bar{N_\mathcal{S}}^{(\varphi_\ell)} + N_\mathcal{S}^{(\varphi_\ell)} - 1\right\}$ directly corresponds to the index i of the CB entries $\mathbf{c}^{(i)} \in \mathcal{C}$ and thus defines the states $s_\ell = i \in \mathcal{S}$ of the HMM. This concludes the 1st step of the ABE training procedure as depicted in Fig. 2.1.

2.1.3 LDA Training (2nd Step)

After the supervised CB training has been performed, a linear transformation matrix $\mathbf{H}$ is trained via LDA in the 2nd step of Fig. 2.1. LDA is a well-known classification method in the field of pattern recognition (Fukunaga, 1990, Chap. 10). It is employed to reduce the dimension of feature vector $\tilde{\mathbf{x}}_\ell$ yielding a feature decorrelation. This relaxes the complexity of the HMM, which is subsequently trained by means of the LDA-transformed feature vectors

$\mathbf{x}_\ell \; \forall \; \ell = 0, 1, \ldots, L-1$.

Feature vector $\tilde{\mathbf{x}}_\ell$ contains static features being extracted from the NB speech data $s_{\mathrm{NB}}(n')$ frame by frame[4]. Furthermore, it includes first- and second-order dynamic features derived from the static ones. On the one hand, a derivation for online applications will be shown spending only one frame of algorithmic delay. On the other hand, a general derivation approach will be demonstrated being able to use more latency, which is feasible for offline applications. Based on the composite feature vectors $\tilde{\mathbf{x}}_\ell \; \forall \; \ell = 0, 1, \ldots, L-1$, a conventional LDA is conducted according to (Jax, 2002, Sec. 5.2). The HMM states $s_\ell = i \in \mathcal{S}$ obtained from the supervised VQ thereby serve as classifications. Thus, the phonetic *a priori* knowledge also influences the LDA training. The finally LDA-transformed feature vector $\mathbf{x}_\ell$ serves as an observation for the subsequent HMM training later in the 3rd step of Fig. 2.1.

Extraction of Static Features

Jax (2002, Sec. 5.4) investigated numerous static features regarding their potential for ABE via the instrumental measures of mutual information and separability. Based on this information-theoretical investigation, he proposed 15 static features (Jax, 2002, Sec. 5.3.5). According to his proposal, the following static features have been selected for this work: Ten normalized ACF coefficients $\tilde{x}_{\mathrm{acf},\ell}(1), \tilde{x}_{\mathrm{acf},\ell}(2), \ldots, \tilde{x}_{\mathrm{acf},\ell}(10)$ as well as the zero-crossing rate $\tilde{x}_{\mathrm{zcr},\ell}$, the gradient index $\tilde{x}_{\mathrm{gi},\ell}$, the spectral centroid $\tilde{x}_{\mathrm{sc},\ell}$, the local kurtosis $\tilde{x}_{\mathrm{lk},\ell}$, and the normalized relative frame energy $\tilde{x}_{\mathrm{rfe},\ell}$. All of these static features are put together into feature vector $\tilde{\mathbf{x}}_\ell^{\mathrm{stat}} = \left[\tilde{x}_{\mathrm{acf},\ell}(1), \tilde{x}_{\mathrm{acf},\ell}(2), \ldots, \tilde{x}_{\mathrm{acf},\ell}(10), \tilde{x}_{\mathrm{zcr},\ell}, \tilde{x}_{\mathrm{gi},\ell}, \tilde{x}_{\mathrm{sc},\ell}, \tilde{x}_{\mathrm{lk},\ell}, \tilde{x}_{\mathrm{rfe},\ell}\right]^{\mathrm{T}} \in \mathbb{R}^{15}$.

The ACF can basically be computed in the frequency domain via (2.9), (2.11), and (2.13). In this work, it is alternatively calculated in the time domain and normalized to its zeroth coefficient representing the NB frame energy $E_\ell = \sum_{n'=0}^{N'_{\mathrm{w}}-1} (s_{\mathrm{NB},\ell}(n'))^2$ (Jax, 2002, Eq. (5.23))

$$\tilde{x}_{\mathrm{acf},\ell}(\nu') = \frac{1}{E_\ell} \sum_{n'=\nu'}^{N'_{\mathrm{w}}-1} s_{\mathrm{NB},\ell}(n'-\nu') s_{\mathrm{NB},\ell}(n'), \quad \nu' = 1, 2, \ldots, 10, \tag{2.19}$$

with ν' denoting the ACF index. The first ACF coefficients are commonly used for LP analysis and therefore contain information about the spectral envelope of the underlying speech frame. Hence, they are adequate features for ABE. The auto-correlation is higher for short-term stationary signals (e.g., voiced speech sounds) than for nonstationary signals (e.g., unvoiced speech sounds).

The short-term average zero-crossing rate is a well-known feature for distinction between voiced and unvoiced speech sounds (Rabiner and Schafer, 1978, Sec. 4.3) and has been widely

[4]In contrast to (Jax, 2002), the feature extraction employs the NB speech data $s_{\mathrm{NB}}(n')$ instead of its interpolated version $s_{\mathrm{LB}}(n)$ to relax the computational complexity particularly for ABE processing in Sec. 2.2.

used for speech recognition (Rabiner and Schafer, 1978, Sec. 9.3). In this work, a normalized variant of the zero-crossing rate is computed according to (Jax, 2002, Eq. (5.16))

$$\tilde{x}_{\mathrm{zcr},\ell} = \frac{1}{N'_{\mathrm{w}} - 1} \sum_{n'=1}^{N'_{\mathrm{w}}-1} \frac{1}{2} \left| \mathrm{sign}\left(s_{\mathrm{NB},\ell}(n'-1)\right) - \mathrm{sign}\left(s_{\mathrm{NB},\ell}(n')\right) \right|, \tag{2.20}$$

with the sign operation being defined as (Jax, 2002, Eq. (5.17))

$$\mathrm{sign}(x) = \begin{cases} +1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ -1, & \text{if } x < 0. \end{cases} \tag{2.21}$$

It represents the number of zero crossings in the underlying speech frame normalized to the maximum possible number $N'_{\mathrm{w}} - 1$. In general, unvoiced speech sounds reveal more zero crossings than voiced speech sounds.

Another voiced/unvoiced classifier is obtained by the gradient index that analyzes changes of the signal direction (Jax, 2002, Sec. 5.15)

$$\tilde{x}_{\mathrm{gi},\ell} = \frac{1}{10\sqrt{E_\ell}} \sum_{n'=1}^{N'_{\mathrm{w}}-1} \left| g_\ell(n') \right| \cdot \frac{1}{2} \left| \mathrm{sign}\left(g_\ell(n')\right) - \mathrm{sign}\left(g_\ell(n'-1)\right) \right|, \tag{2.22}$$

with $g_\ell(n') = s_{\mathrm{NB},\ell}(n') - s_{\mathrm{NB},\ell}(n'-1)$ denoting the signal gradient. In unvoiced speech sounds there are generally more changes of the signal direction than in voiced speech sounds.

The spectral centroid commonly stays low for voiced speech sounds and increases for unvoiced speech sounds. It is computed in the frequency domain via a short-term DFT (Oppenheim and Schafer, 1989, Sec. 8.1)

$$S_{\mathrm{NB},\ell}(k') = \sum_{n'=0}^{K_{\mathrm{sc}}-1} s_{\mathrm{NB},\ell}(n') \cdot e^{-j2\pi \frac{n'k'}{K_{\mathrm{sc}}}}, \quad k' = 0, 1, \ldots, K_{\mathrm{sc}} - 1, \tag{2.23}$$

which involves a zero-padding of the NB speech frame in contrast to (2.9). This allows for a DFT length of $K_{\mathrm{sc}} > N'_{\mathrm{w}}$ being set to $K_{\mathrm{sc}} = 2^{\lceil \log_2(N'_{\mathrm{w}}) \rceil}$, i.e., to the smallest power of two larger than N'_{w}. By means of the first $K_{\mathrm{sc}}/2 + 1$ non-redundant DFT coefficients, a normalized version of the spectral centroid is obtained (Jax, 2002, Eq. (5.20))

$$\tilde{x}_{\mathrm{sc},\ell} = \frac{\sum_{k'=0}^{K_{\mathrm{sc}}/2} k' \cdot \left| S_{\mathrm{NB},\ell}(k') \right|}{\left(\frac{K_{\mathrm{sc}}}{2} + 1\right) \sum_{k'=0}^{K_{\mathrm{sc}}/2} \left| S_{\mathrm{NB},\ell}(k') \right|}. \tag{2.24}$$

Please note that the spectral centroid represents an important feature to detect relevant phonemes for ABE according to Sec. 1.3, such as the critical fricatives /s/ and /z/.

The local kurtosis is a short-term estimate of the kurtosis measure indicating how normally distributed a signal is. Originating from higher-order statistics, it involves the forth- and second-order moments of a speech frame (Krishnamachari et al., 2001)

$$E\left\{(s_{\mathrm{NB},\ell}(n'))^4\right\} = \frac{1}{N'_{\mathrm{w}}} \sum_{n'=0}^{N'_{\mathrm{w}}-1} (s_{\mathrm{NB},\ell}(n'))^4, \tag{2.25}$$

$$E\left\{(s_{\mathrm{NB},\ell}(n'))^2\right\} = \frac{1}{N'_{\mathrm{w}}} \sum_{n'=0}^{N'_{\mathrm{w}}-1} (s_{\mathrm{NB},\ell}(n'))^2. \tag{2.26}$$

They are put into relation via the common logarithm (Jax, 2002, Eq. (5.19))

$$\tilde{x}_{\mathrm{lk},\ell} = \log_{10}\left(\frac{E\left\{(s_{\mathrm{NB},\ell}(n'))^4\right\}}{\left(E\left\{(s_{\mathrm{NB},\ell}(n'))^2\right\}\right)^2}\right). \tag{2.27}$$

Most of the voiced sounds reveal a local kurtosis of less than $\log_{10}(3)$, however, for onsets of strong vowels and of plosives it usually increases significantly (Jax, 2002).

The frame energy E_ℓ is able to reliably distinguish voiced from unvoiced speech sounds and to detect speech activity at least for a sufficiently high signal-to-noise ratio (SNR) (Rabiner and Schafer, 1978, Sec. 4.2). To make this feature more robust against background noise, the noise floor $E_{\mathrm{min},\ell}$ can be eliminated. It is estimated as the minimum frame energy of the last L_{min} frames[5] (Jax, 2002, Eq. (5.26))

$$E_{\mathrm{min},\ell} = \min_{l\in\{0,\ldots,L_{\mathrm{min}}-1\}} E_{\ell-l}, \tag{2.28}$$

with L_{min} being set to $\left\lfloor \frac{t_{\mathrm{min}}}{N'/f'_{\mathrm{s}}} \right\rceil$ and $t_{\mathrm{min}} = 0.625$ s denoting the time span in which the noise floor can be assumed to be quasi-stationary (Martin, 1993, 2001). Furthermore, an average frame energy $\bar{E}_\ell$ can be used for normalization purposes, to reduce dependencies on long-term speech level variations. It is determined by smoothing E_ℓ via a first-order infinite impulse response (IIR) recursion[6] (Jax, 2002, Eq. 5.25)

$$\bar{E}_\ell = \alpha\bar{E}_{\ell-1} + (1-\alpha)E_\ell, \tag{2.29}$$

with factor $\alpha = 0.96$ being set close to one. By means of $E_{\mathrm{min},\ell}$ and $\bar{E}_\ell$, the normalized relative frame energy is finally formulated as a logarithmic expression (Jax, 2002, Eq. 5.29)

$$\tilde{x}_{\mathrm{rfe},\ell} = \frac{\log_{10}(E_\ell) - \log_{10}(E_{\mathrm{min},\ell})}{\log_{10}(\bar{E}_\ell) - \log_{10}(E_{\mathrm{min},\ell})}. \tag{2.30}$$

[5]The frame energy needs to be initialized appropriately for the noise floor estimation of the first $L_{\mathrm{min}}-1$ frames, e.g., by a large positive value.

[6]The average frame energy needs to be initialized appropriately for the IIR recursion of the zeroth frame, e.g., by taking the mean of the frame energy over all NB training speech data.

Derivation of Dynamic Features

To better capture time dependencies, the first- and second-order dynamic feature vectors $\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}}$ and $\Delta\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}}$, respectively, are derived from the static features. In literature, they are commonly known as Δ- and $\Delta\Delta$-features (Huang et al., 2001, Sec. 9.3.3). Their purpose is to improve the detection of speech sounds with a short duration, such as plosives and fricatives. However, a derivation in time requires the static feature vector $\tilde{\mathbf{x}}_\ell^{\text{stat}}$ to be available for future (and past) frames, which provokes an additional algorithmic delay. Depending on the application, more or less algorithmic delay can be acceptable. This work focuses on two ABE applications. On the one hand, Chap. 3 deals with an *offline* ABE for training WB telephony ASR systems. On the other hand, an *online* ABE for enhancement of NB telephone speech services is treated in Chap. 4. Hence, two different methods for dynamic feature derivation are presented: While the first one involves an algorithmic delay of five frames and is therefore only acceptable for offline ABE, the second method introducing a latency of just one frame is suited to online ABE.

The first derivation method includes polynomial approximations of the static and first-order dynamic feature trajectory (Young et al., 2006, Sec. 5.9)

$$\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}} = \frac{\sum_{\theta=1}^{\Theta_\Delta} \theta \cdot \left(\tilde{\mathbf{x}}_{\ell+\theta}^{\text{stat}} - \tilde{\mathbf{x}}_{\ell-\theta}^{\text{stat}}\right)}{2\sum_{\theta=1}^{\Theta_\Delta} \theta^2} = \frac{\sum_{\theta=-\Theta_\Delta}^{\Theta_\Delta} \theta \cdot \tilde{\mathbf{x}}_{\ell+\theta}^{\text{stat}}}{\sum_{\theta=-\Theta_\Delta}^{\Theta_\Delta} \theta^2}, \tag{2.31}$$

$$\Delta\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}} = \frac{\sum_{\theta=1}^{\Theta_{\Delta\Delta}} \theta \cdot \left(\Delta\tilde{\mathbf{x}}_{\ell+\theta}^{\text{stat}} - \Delta\tilde{\mathbf{x}}_{\ell-\theta}^{\text{stat}}\right)}{2\sum_{\theta=1}^{\Theta_{\Delta\Delta}} \theta^2} = \frac{\sum_{\theta=-\Theta_{\Delta\Delta}}^{\Theta_{\Delta\Delta}} \theta \cdot \Delta\tilde{\mathbf{x}}_{\ell+\theta}^{\text{stat}}}{\sum_{\theta=-\Theta_{\Delta\Delta}}^{\Theta_{\Delta\Delta}} \theta^2}, \tag{2.32}$$

respectively. The polynomial orders $\Theta_\Delta = 3$ and $\Theta_{\Delta\Delta} = 2$ denote the respective number of future/past frames to be taken into account (Bauer et al., 2014b, Eq. (1)). This results in a total algorithmic delay of five frames (i.e., $\Theta_\Delta + \Theta_{\Delta\Delta} = 5$).

In contrast, the second derivation method represents simple difference equations, which are just based on the static feature trajectory and only require a latency of one frame (Bauer et al., 2010b)

$$\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}} = \tilde{\mathbf{x}}_{\ell+1}^{\text{stat}} - \tilde{\mathbf{x}}_{\ell-1}^{\text{stat}}, \tag{2.33}$$

$$\Delta\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}} = \left(\tilde{\mathbf{x}}_{\ell+1}^{\text{stat}} - \tilde{\mathbf{x}}_{\ell}^{\text{stat}}\right) - \left(\tilde{\mathbf{x}}_{\ell}^{\text{stat}} - \tilde{\mathbf{x}}_{\ell-1}^{\text{stat}}\right) = \tilde{\mathbf{x}}_{\ell+1}^{\text{stat}} - 2\tilde{\mathbf{x}}_{\ell}^{\text{stat}} + \tilde{\mathbf{x}}_{\ell-1}^{\text{stat}}. \tag{2.34}$$

Finally, all static and dynamic features are put together into a composite feature vector

of dimension $\tilde{d} = 45$:

$$\tilde{\mathbf{x}}_\ell = \begin{bmatrix} \tilde{\mathbf{x}}_\ell^{\text{stat}} \\ \Delta\tilde{\mathbf{x}}_\ell^{\text{stat}} \\ \Delta\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}} \end{bmatrix} \in \mathbb{R}^{\tilde{d}}. \tag{2.35}$$

LDA of Feature Vectors and Transformation

Instead of directly using the composite feature vectors $\tilde{\mathbf{x}}_\ell$ as observations for the subsequent HMM training, an LDA is employed in Fig. 2.1 to relax the complexity of the statistical model. The LDA is a well-known classification method from the field of pattern recognition (Fukunaga, 1990, Chap. 10). It aims at reducing the dimension of feature vectors by retaining their discriminating power to a great extent. Furthermore, the elements of the LDA-transformed feature vectors are being mutually decorrelated. This allows for the use of diagonal instead of full covariance matrices in Sec. 2.1.4.

In this work, the LDA is implemented according to (Jax, 2002, Sec. 5.2). By means of a $\tilde{d} \times d$ matrix $\mathbf{H}$, the composite feature vector $\tilde{\mathbf{x}}_\ell \in \mathbb{R}^{\tilde{d}}$ in (2.35) is linearly transformed into a feature vector $\mathbf{x}_\ell \in \mathbb{R}^d$ of reduced dimension $d < \tilde{d}$ by

$$\mathbf{x}_\ell = \mathbf{H}^{\mathrm{T}}\tilde{\mathbf{x}}_\ell. \tag{2.36}$$

The linear transformation matrix $\mathbf{H}$ is trained based on the composite feature vectors $\tilde{\mathbf{x}}_\ell \ \forall \ \ell = 0, 1, \ldots, L-1$ using the HMM states $s_\ell = i \in \mathcal{S}$ as class labels. Due to the involved eigenvalue problem, the reduced feature dimension d needs to be smaller than the number of HMM states $N_\mathcal{S}$ (Jax, 2002, Sec. 5.2.2), i.e., $d \in \left[1, \min\{\tilde{d}, N_\mathcal{S}\}\right)$.

The training of $\mathbf{H}$ underlies the maximization of the separability function $\varsigma_\mathbf{x}(d)$ (Jax, 2002, Eq. (5.13)). It is defined as the sum of the d largest eigenvalues that derive from the empirical separability matrix $\mathbf{J}_{\tilde{\mathbf{x}}} = {\mathbf{W}_{\tilde{\mathbf{x}}}}^{-1}\mathbf{B}_{\tilde{\mathbf{x}}}$, with $\mathbf{W}_{\tilde{\mathbf{x}}}$ and $\mathbf{B}_{\tilde{\mathbf{x}}}$ denoting the within- and between-class covariance matrices of the non-transformed feature vectors (Jax, 2002, Sec. 5.2.1), respectively. This criterion assumes that the feature vectors assigned to a class are normally distributed. By adding up the diagonal elements of $\mathbf{J}_{\tilde{\mathbf{x}}}$, the original separability of the feature vectors before LDA $\varsigma_{\tilde{\mathbf{x}}}$ can be obtained (Jax, 2002, Eq. (5.6)). It represents the maximum achievable value of $\varsigma_\mathbf{x}(d)$. The reduced feature dimension d should be high enough, so that the LDA can preserve most of the original separability. However, a higher separability does not necessarily lead to a better ABE performance.

Based on all 5578 phonetically rich sentences provided by the close-talk recordings of the American English speech corpus SpeechDat-Car US (Moreno et al., 2000), the non-transformed composite feature vectors $\tilde{\mathbf{x}}_\ell$ with 'low-delay' derivatives (2.33)–(2.34) interestingly reveal a higher value of $\varsigma_{\tilde{\mathbf{x}}}$ than those including the dynamic features for offline ABE (2.31)–(2.32). As expected, the lowest separability before LDA is attained by the static

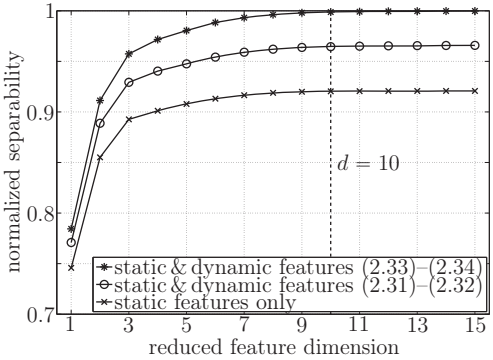


Figure 2.6: Separabilities after LDA normalized to the highest separability before LDA.

feature vectors $\tilde{\mathbf{x}}_\ell^{\text{stat}}$. Fig. 2.6 illustrates the separability function $\varsigma_{\mathbf{x}}(d)$ for the transformed feature vectors with and without dynamic features. For normalization purposes, all curves are divided by the highest value of $\varsigma_{\tilde{\mathbf{x}}}$. Obviously, they increase monotonically with the reduced feature dimension and are sufficiently saturated at $d = 10$. Hence, this value is used for the LDA in the context of ABE throughout this work. This concludes the 2nd step of the ABE training procedure as depicted in Fig. 2.1.

2.1.4 HMM Training (3rd Step)

Now let us focus on the 3rd step in Fig. 2.1, the HMM training. HMMs are commonly used for statistical modeling of time-varying random processes, e.g., in the field of speech recognition, and can be widely found in literature (Rabiner, 1989; Rabiner and Juang, 1993; Huang et al., 2001). This work employs for the purpose of ABE a first-order[7] HMM following (Jax, 2002, Sec. 2.2 and 6.3). Its states $s_\ell \in \mathcal{S}$ are defined within training via the supervised VQ in Sec. 2.1.2. Thus, they are influenced by phonetic *a priori* knowledge in terms of the phoneme class labels φ_ℓ, as opposed to (Jax, 2002, Sec. 6.2).

Based on the LDA-transformed feature vectors $\mathbf{x}_\ell$ from Sec. 2.1.3 and the predefined HMM states $s_\ell = i$ from Sec. 2.1.2, the training of the HMM parameters takes place, as depicted in Fig. 2.1. On the one hand, the initial state probabilities $\pi_i = \mathrm{P}(s_0 = i)\ \forall\ i \in \mathcal{S}$ as well as the elements of the state transition probability matrix $a_{i,j} = \mathrm{P}(s_\ell = j | s_{\ell-1} = i)\ \forall\ i, j \in \mathcal{S}$ are calculated by just counting the state occurrences. On the other hand, the state observation likelihoods $b_j(\mathbf{x}_\ell) = \mathrm{p}(\mathbf{x}_\ell | s_\ell = j)\ \forall\ j \in \mathcal{S}$ are approximated by GMMs, which are iteratively trained via the well-known expectation maximization (EM) algorithm (Dempster et al., 1977). The complete set of HMM parameters is given by $\lambda = \{\pi_i, a_{i,j}, b_j(\mathbf{x}) \mid i, j \in \mathcal{S}\}$.

[7] In a Markov chain of order one, any state only depends on its predecessor.

Initial State and State Transition Probabilities

The Markov chain is characterized by the elements of the state transition probability matrix $a_{i,j} = \mathrm{P}(s_\ell = j | s_{\ell-1} = i)\ \forall\ i, j \in \mathcal{S}$. To calculate them as well as the initial state probabilities $\pi_i = \mathrm{P}(s_0 = i)\ \forall\ i \in \mathcal{S}$, the two-dimensional joint state histogram $\mathrm{H}(s_\ell = j, s_{\ell-1} = i)\ \forall\ i, j \in \mathcal{S}$ is required. It is derived by counting the numbers of transitions from state $s_{\ell-1} = i$ to state $s_\ell = j$ for $\ell = 1, 2, \ldots, L-1$. When normalizing the resulting counts by the total number of state transitions, joint state probabilities are obtained

$$\mathrm{P}(s_\ell = j, s_{\ell-1} = i) = \frac{\mathrm{H}(s_\ell = j, s_{\ell-1} = i)}{L-1} \quad \forall\ i, j \in \mathcal{S}, \tag{2.37}$$

which meet the stochastic constraint $\sum_{j=0}^{N_\mathcal{S}-1} \sum_{i=0}^{N_\mathcal{S}-1} \mathrm{P}(s_\ell = j, s_{\ell-1} = i) = 1$. Based on these, the initial state and state transition probabilities are computed as (Rabiner, 1989)

$$\pi_i = \sum_{j=0}^{N_\mathcal{S}-1} \mathrm{P}(s_1 = j, s_0 = i) \quad \forall\ i \in \mathcal{S}, \tag{2.38}$$

$$a_{i,j} = \frac{\mathrm{P}(s_\ell = j, s_{\ell-1} = i)}{\pi_i} \quad \forall\ i, j \in \mathcal{S}, \tag{2.39}$$

respectively. Again, the stochastic constraints $\sum_{i=0}^{N_\mathcal{S}-1} \pi_i = 1$ and $\sum_{j=0}^{N_\mathcal{S}-1} a_{i,j} = 1\ \forall\ i \in \mathcal{S}$ hold.

Ergodic HMMs have the property that basically any state can be followed by any other one, i.e., $a_{i,j} > 0\ \forall\ i, j \in \mathcal{S}$ (Rabiner and Juang, 1993, Sec. 6.3.3). Due to insufficient training data (Rabiner and Juang, 1993, Sec. 6.12.4), however, some elements of the state transition probability matrix may turn out to be zero. This kind of sparse data problem is tackled in Sec. 3.3.1 by introducing a state transition smoothing (Huang et al., 2001, Sec. 8.4.5).

State Observation Likelihoods

The hidden states of an HMM are not directly observable. Hence, the feature vectors $\mathbf{x}_\ell$ are used for an indirect state observation. As they are neither discrete nor vector-quantized events in the present case, the state observation likelihoods $b_j(\mathbf{x}_\ell)\ \forall\ j \in \mathcal{S}$ need to be modeled by continuous probability density functions (PDFs) characterizing a continuous-density HMM (Rabiner, 1989; Rabiner and Juang, 1993; Huang et al., 2001). A well-tried choice is to approximate them by GMMs (Huang et al., 2001, Sec. 8.3.1). For this purpose, a particular GMM is dedicated to each state $j \in \mathcal{S}$ according to (Jax, 2002, Sec. 6.3.1). Thus, the state-conditional observation likelihoods are modeled by a weighted sum of multivariate Gaussian PDFs (Reynolds and Rose, 1995)

$$b_j(\mathbf{x}_\ell) \approx \sum_{m=0}^{M-1} \rho_{j,m}\, \mathcal{N}\left(\mathbf{x}_\ell; \boldsymbol{\mu}_{j,m}, \boldsymbol{\Sigma}_{j,m}\right), \tag{2.40}$$

with the GMM parameters of state j and mixture component m being denoted by mixture weight $\rho_{j,m}$, mean vector $\boldsymbol{\mu}_{j,m}$ and covariance matrix $\boldsymbol{\Sigma}_{j,m}$. The mixture weights $\rho_{j,m} \in [0,1]$ thereby satisfy the stochastic constraint $\sum_{m=0}^{M-1} \rho_{j,m} = 1 \ \forall \ j \in \mathcal{S}$ (Huang et al., 2001, Sec. 8.3.1). Each mixture density is represented by a d-dimensional normal distribution (Reynolds and Rose, 1995, Eq. (2))

$$\mathcal{N}\left(\mathbf{x}_\ell ; \boldsymbol{\mu}_{j,m}, \boldsymbol{\Sigma}_{j,m}\right) = \frac{1}{(2\pi)^{d/2} \ (\det \boldsymbol{\Sigma}_{j,m})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_\ell - \boldsymbol{\mu}_{j,m})^{\mathrm{T}} \ \boldsymbol{\Sigma}_{j,m}^{-1} \ (\mathbf{x}_\ell - \boldsymbol{\mu}_{j,m})\right). \quad (2.41)$$

A proper value for the GMM order, i.e., the number of mixture components, leading to a good approximation of $b_j(\mathbf{x}_\ell)$ was found to be $M = 8$ (Jax, 2002, Fig. 4.8).

The most commonly used way of GMM training is the maximum likelihood (ML) parameter estimation, which aims at finding those GMM parameters that maximize the modeled likelihood (Reynolds and Rose, 1995, Sec. II.D). It can be iteratively conducted via the well-known EM algorithm (Dempster et al., 1977). Due to the strictly monotonic increase of the likelihood from iteration to iteration, the EM algorithm at least approaches a local maximum, even though a convergence to the global maximum cannot be assured (Dempster et al., 1977). The EM algorithm employed in this work terminates, if the relative log-likelihood increase among two iterations gets smaller than a predefined threshold of 10^{-6} (Jax, 2002, Eq. (4.47)). This stop condition turned out to be more robust than, e.g., the termination after a fixed number of iterations.

Before applying the EM algorithm, the LDA-transformed feature vectors $\mathbf{x}_\ell$ need to be assigned by means of the HMM states $s_\ell \in \mathcal{S}$ into $N_\mathcal{S}$ clusters, according to (Jax, 2002, Sec. 4.5.1). For each of these state-specific clusters, a separate GMM is trained. In order to initialize the separately applied EM algorithms, all clusters need to be further subdivided into M sub-clusters. In principle, this mixture-specific clustering could be done randomly, as the EM initialization plays an inferior role (Reynolds and Rose, 1995, Sec. III.C1). In this work, however, an LBG-based VQ of the UB cepstral envelope vectors $\mathbf{c}_\ell$ is used for initialization purposes[8], according to (Jax, 2002, Sec. 4.5.1).

Due to the de-correlation of the feature vectors $\mathbf{x}_\ell$ by means of the LDA transformation in advance of the GMM training, diagonal instead of full covariance matrices are used. This relaxes the complexity of the statistical model and also tackles the problem that covariance matrices may become singular (i.e., not invertible) due to insufficient training data (Huang et al., 2001, Sec. 8.4.5). If singularities still appear in spite of the diagonal structure of $\boldsymbol{\Sigma}_{j,m}$, they are removed within each EM iteration step by means of a variance limiting (Reynolds and Rose, 1995, Sec. III.C1). A variance floor of 10^{-5} thereby turned out to be large enough. This concludes the 3$^{\mathrm{rd}}$ step of the ABE training procedure as depicted in Fig. 2.1.

[8]To simplify matters, a connection between the cepstral UB calculation block and the EM algorithm block is not visualized in Fig. 2.1.

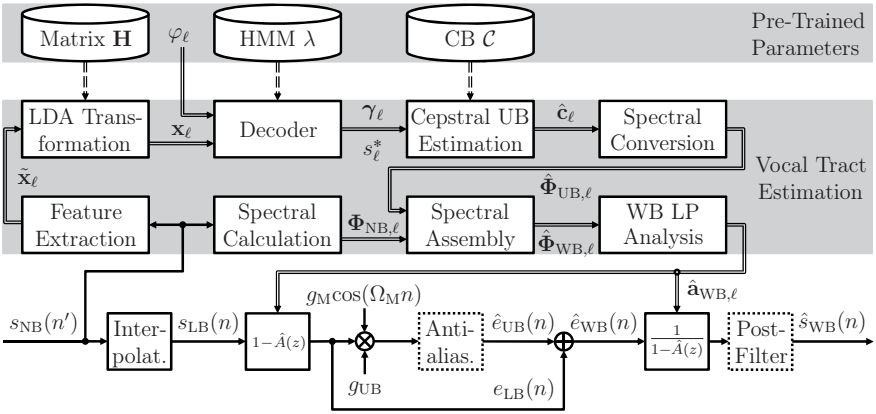


Figure 2.7: Block diagram of ABE processing exploiting phonetic *a priori* knowledge.

2.2 ABE Processing with Phonetic Support

After the ABE parameters implying the CB $\mathcal{C}$, the LDA matrix $\mathbf{H}$, and the HMM λ have been trained, the actual ABE processing can be performed. It relies on the algorithmic concept in (Jax, 2002, Sec. 2.3), however, some crucial steps of this reference algorithm are developed further. In particular, the exploitation of phonetic *a priori* knowledge corresponding to the ABE training in Sec. 2.1 represents the most important innovation. The proposed ABE processing is divided into two steps, as can be seen in Fig. 2.7. On the one hand, the main ABE processing is conducted in the lower signal path. On the other hand, a vocal tract estimation is performed in the upper signal paths by means of the pre-trained parameters including the prepared phoneme class labels[9] φ_ℓ.

According to Fig. 2.1, the signaling paths in Fig. 2.7 are characterized by single and double lines to differentiate between a sample- and frame-wise processing. Obviously, the main ABE processing is done sample by sample, whereas the vocal tract is estimated frame by frame. The NB input speech samples $s_{\mathrm{NB}}(n')$ therefore need to be directly converted into frames within the feature extraction and spectral calculation blocks. For this purpose, a frame conversion is applied taking into account 50 % symmetrically overlapping Blackman windows, as specified in Sec. 2.1.1. The way back into the sample-wise processing takes place at the WB LP analysis and synthesis filtering blocks $1-\hat{A}(z)$ and $(1-\hat{A}(z))^{-1}$, respectively. There, the estimated WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$ are switched every frame $\ell = 0, 1, \ldots, L-1$ at

[9]For convenience, the nomenclature of φ_ℓ does not differentiate between ABE training and processing throughout this work, although disjoint data sets are used. This also applies to other variables utilized twice, such as $s_{\mathrm{NB}}(n')$, $s_{\mathrm{NB},\ell}(n')$, $\Phi_{\mathrm{NB},\ell}(k')$, $\tilde{\mathbf{x}}_\ell$, $\mathbf{x}_\ell$, s_ℓ, $\sigma_{\mathrm{LB},\ell}$, and L.

sample index $n = \ell \cdot N_\mathrm{s}$.

Main ABE Processing

By interpolating the NB input speech $s_\mathrm{NB}(n')$ to a sampling rate of 16 kHz, an interpolated LB speech $s_\mathrm{LB}(n)$ is obtained. It is subject to the finite impulse response (FIR) WB LP analysis filter $1-\hat{A}(z)$ modeling the spectral envelope of the inverse vocal tract. A time-domain modulation of the resulting LB residual signal $e_\mathrm{LB}(n)$ yields an estimated UB residual signal $\hat{e}_\mathrm{UB}(n)$. Depending on the modulation frequency Ω_M, an IIR highpass filter may be required to remove aliasing in the LB frequency range. In contrast to the time-domain modulations in (Jax, 2002, Sec. 3.3), a static weight $g_\mathrm{UB} \in (0, 1]$ is used to attenuate the UB frequencies in the residual domain. In this way, the aggressiveness of the ABE can be controlled. A superposition of $e_\mathrm{LB}(n)$ and $\hat{e}_\mathrm{UB}(n)$ results in the extended residual signal $\hat{e}_\mathrm{WB}(n)$ serving as a WB excitation of the vocal tract. Within the next step, $\hat{e}_\mathrm{WB}(n)$ is spectrally shaped via the autoregressive WB LP synthesis filter $(1-\hat{A}(z))^{-1}$ modeling the spectral envelope of the vocal tract. In addition to (Jax, 2002, Sec. 2.3), the upper cut-off frequency of the estimated WB output speech $\hat{s}_\mathrm{WB}(n)$ can finally be adapted by means of an optional IIR lowpass postfilter to control the degree of bandwidth extension.

Please note that the transparency of the LB spectrum due to the inverse filters $1-\hat{A}(z)$ and $(1-\hat{A}(z))^{-1}$ is still preserved. This ensures that the LB frequencies remain unchanged, which represents – as discussed in (Jax, 2002, Sec. 2.4) – an important ABE property.

Vocal Tract Estimation

The required WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$ are estimated in the upper signal paths. To reduce computational complexity, the feature extraction and spectral calculation blocks directly access the NB input speech $s_\mathrm{NB}(n')$ in spite of its interpolated version $s_\mathrm{LB}(n)$, as opposed to (Jax, 2002, Sec. 3.3). For the sake of consistency with the ABE training, the extraction of the composite feature vector $\tilde{\mathbf{x}}_\ell$ as well as the subsequent LDA transformation using the pre-trained LDA matrix $\mathbf{H}$ are conducted according to Sec. 2.1.3.

Based on the LDA-transformed feature vector $\mathbf{x}_\ell$, the pre-trained HMM λ is evaluated to calculate *a posteriori* probabilities $\boldsymbol{\gamma}_\ell$. In contrast to (Jax, 2002, Sec. 6.4.1), phonetic *a priori* knowledge is thereby involved in terms of the phoneme class labels φ_ℓ. This requires an adaptation of the employed optimal state decoder, based on the well-known forward algorithm (FA) or forward-backward algorithm (FBA) (Rabiner and Juang, 1993, Sec. 6.4.1), according to (Bauer et al., 2014b). Furthermore, the HMM evaluation is augmented with the commonly used Viterbi algorithm (VA) (Rabiner and Juang, 1993, Sec. 6.4.2), in order to decode the optimal state sequence $(s^*)_0^{L-1} = (s_0^*, s_1^*, \ldots, s_\ell^*, \ldots, s_{L-1}^*)$ following (Bauer et al.,

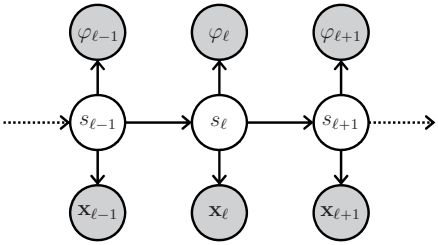


Figure 2.8: HMM dependency scheme including both feature and phonetic observations.

2014b). The use of phonetic *a priori* knowledge for decoding the optimal state sequence aims at improving the performance of the HMM-based estimation process, particularly regarding the ABE-relevant phonemes identified in Sec. 1.3.

Taking into account $\boldsymbol{\gamma}_\ell$ and s_ℓ^*, different estimation rules can be used to obtain an UB cepstral envelope estimate $\hat{\mathbf{c}}_\ell$ from the pre-trained CB $\mathcal{C}$. The subsequent spectral conversion into the estimated UB short-term power spectrum $\hat{\boldsymbol{\Phi}}_{\mathrm{UB},\ell}$ is conducted according to (Jax, 2002, Eq. (6.3)), however, under a moderate cepstral interframe smoothing constraint (Bauer and Fingscheidt, 2009a,b). This prevents in advance potential switching effects of the time-variant WB LP filter coefficients (Schnell and Lacroix, 2008). To be consistent with the ABE training, the NB short-term power spectrum $\boldsymbol{\Phi}_{\mathrm{NB},\ell}$ is calculated via (2.9) and (2.11). By means of $\hat{\boldsymbol{\Phi}}_{\mathrm{UB},\ell}$ and $\boldsymbol{\Phi}_{\mathrm{NB},\ell}$, the estimated WB short-term power spectrum $\hat{\boldsymbol{\Phi}}_{\mathrm{WB},\ell}$ is spectrally assembled following (Jax, 2002, Eq. (6.5)). The LB prediction gain factor $\sigma_{\mathrm{LB},\ell}$ is thereby used to estimate the energy of the extended frequency band. It is acquired via a separate LB LP analysis corresponding to the upper signal path in Fig. 2.3. Subsequently, another LP analysis is performed, but this time based on $\hat{\boldsymbol{\Phi}}_{\mathrm{WB},\ell}$ to obtain the estimated WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$.

2.2.1 Phonetic Weighting of Observation Likelihoods

Based on the assumption that the phoneme class labels φ_ℓ in Fig. 2.7 are available for ABE processing, they serve as additional observations along with the feature vectors $\mathbf{x}_\ell$. Some experiments have been carried out by simply concatenating $\tilde{\mathbf{x}}_\ell$ and φ_ℓ as a composed feature vector before LDA. However, the resulting ABE performance could not be noticeably improved by this early integration of phonetic information. As recommended by Lucey et al. (2005), a later stage of information integration – namely at the HMM decoder – is therefore used in this work. The resulting HMM dependency scheme depicted in Fig. 2.8 shows that both feature and phonetic observations are emitted by the hidden states s_ℓ. Thus, the GMM-based state observation likelihood $b_j(\mathbf{x}_\ell) = \mathrm{p}(\mathbf{x}_\ell | s_\ell = j)$ needs to be modified. Obviously, this

modification leads to a joint state observation likelihood $b_j(\mathbf{x}_\ell, \varphi_\ell) = \mathrm{p}(\mathbf{x}_\ell, \varphi_\ell | s_\ell = j)$.

In principle, a GMM could be trained modeling this conditional joint observation PDF. However, linear correlations between $\mathbf{x}_\ell$ and φ_ℓ should then have to be taken into account requiring the use of full instead of diagonal covariance matrices (Jax, 2002, Sec. 6.3.2). Thus, sparse data problems could be caused in case of insufficient training data (Huang et al., 2001, Sec. 8.4.5). Furthermore, the complexity of the statistical model would increase significantly.

For these reasons, the joint observation PDF is split into individual terms of $\mathbf{x}_\ell$ and φ_ℓ by means of the chain rule (Papoulis and Pillai, 2002, Sec. 7.2)

$$\begin{aligned} b_j(\mathbf{x}_\ell, \varphi_\ell) &= \mathrm{p}(\mathbf{x}_\ell, \varphi_\ell | s_\ell = j) \\ &= \frac{\mathrm{p}(\mathbf{x}_\ell, \varphi_\ell, s_\ell = j)}{\mathrm{P}(s_\ell = j)} \\ &= \frac{\mathrm{p}(\mathbf{x}_\ell | \varphi_\ell, s_\ell = j)\, \mathrm{P}(\varphi_\ell | s_\ell = j)\, \mathrm{P}(s_\ell = j)}{\mathrm{P}(s_\ell = j)} \\ &= \mathrm{p}(\mathbf{x}_\ell | \varphi_\ell, s_\ell = j)\, \mathrm{P}(\varphi_\ell | s_\ell = j) \\ &\approx \mathrm{p}(\mathbf{x}_\ell | s_\ell = j)\, \mathrm{P}(\varphi_\ell | s_\ell = j) \\ &= b_j(\mathbf{x}_\ell)\, \mathrm{P}(\varphi_\ell | s_\ell = j). \end{aligned} \tag{2.42}$$

Due to the tight relation between states and phoneme classes illustrated in Fig. 2.5, the expression $\mathrm{p}(\mathbf{x}_\ell | \varphi_\ell, s_\ell = j)$ can be simplified to $\mathrm{p}(\mathbf{x}_\ell | s_\ell = j) = b_j(\mathbf{x}_\ell)$ (Bauer and Fingscheidt, 2009a). As a result, the joint state likelihood $b_j(\mathbf{x}_\ell, \varphi_\ell)$ is decomposed into the original state observation likelihood $b_j(\mathbf{x}_\ell)$, which depends on the feature observation $\mathbf{x}_\ell$, and a conditional probability $\mathrm{P}(\varphi_\ell | s_\ell = j)$ depending on the phonetic observation φ_ℓ. The latter term $\mathrm{P}(\varphi_\ell | s_\ell = j)$ denotes the elements of an $N_\mathcal{P} \times N_\mathcal{S}$ phoneme class probability matrix (Bauer et al., 2014b). Following (Bauer and Fingscheidt, 2009a), the phoneme class probabilities serve as phonetic weights and are defined as

$$\mathrm{P}(\varphi_\ell | s_\ell = j) = \begin{cases} 1 - \varepsilon^{(j)}, & \text{if } N_\mathcal{P} > 1,\ j \in \mathcal{S}^{(\varphi_\ell)}, \\ \frac{\varepsilon^{(j)}}{N_\mathcal{P} - 1}, & \text{if } N_\mathcal{P} > 1,\ j \notin \mathcal{S}^{(\varphi_\ell)}, \\ 1, & \text{else}, \end{cases} \tag{2.43}$$

with $\varepsilon^{(j)}\ \forall\ j \in \mathcal{S}$ being small state-specific values. Please note that a phonetic weighting is only reasonable for more than one phoneme class. Hence, the following discussion is based on the assumption that $N_\mathcal{P} > 1$.

If the observed phonetic observation φ_ℓ and the given HMM state $s_\ell = j$ fit together, i.e., $j \in \mathcal{S}^{(\varphi_\ell)}$, the phoneme class probability is generally much higher (as expressed by the upper weight $1 - \varepsilon^{(j)}$ in (2.43)) than in case of a mismatch (as expressed by the middle weight $\frac{\varepsilon^{(j)}}{N_\mathcal{P} - 1}$ in (2.43)). Thus, $b_j(\mathbf{x}_\ell)$ is phonetically weighted by $\mathrm{P}(\varphi_\ell | s_\ell = j)$. The phonetic weighting is

driven by the small state-specific values

$$\varepsilon^{(j)} \in \left(0, \frac{N_{\mathcal{P}}-1}{N_{\mathcal{P}}}\right] \ \forall\ j \in \mathcal{S}. \tag{2.44}$$

They express a reliability of the underlying phonetic transcription process that has been done either automatically by a forced Viterbi alignment or manually by humans. On the one hand, values of $\varepsilon^{(j)}\ \forall\ j \in \mathcal{S}$ close to zero indicate reliable phoneme class labels and therefore allow for a strong phonetic weighting. On the other hand, $\varepsilon^{(j)} = \varepsilon = \frac{N_{\mathcal{P}}-1}{N_{\mathcal{P}}}\ \forall\ j \in \mathcal{S}$ means that all elements of the phoneme class probability matrix result in $\frac{1}{N_{\mathcal{P}}}$ yielding a uniform distribution, i.e., there is no influence of φ_ℓ. Independent from the choice of $\varepsilon^{(j)}$, the stochastic constraint $\sum_{\varphi\in\mathcal{P}} \mathrm{P}(\varphi|s_\ell = j) = 1\ \forall\ j \in \mathcal{S}$ is fulfilled anyway.

An adequate parametrization of $\varepsilon^{(j)}$, which also takes into account the number of phoneme classes $N_{\mathcal{P}} > 1$, can be determined via

$$\varepsilon^{(j)} = \frac{1}{1 + \frac{r^{(j)}}{N_{\mathcal{P}}-1}}\ \forall\ j \in \mathcal{S}, \tag{2.45}$$

with $r^{(j)}$ defining the ratio between the upper and middle phonetic weight in (2.43). Please note that (2.45) is based on the definition

$$r^{(j)} = \frac{1-\varepsilon^{(j)}}{\frac{\varepsilon^{(j)}}{N_{\mathcal{P}}-1}}\ \forall\ j \in \mathcal{S}. \tag{2.46}$$

Assuming $N_{\mathcal{P}} > 1$, $r^{(j)}$ can take on values within $[1, \infty)$ corresponding to the range of $\varepsilon^{(j)}$ defined in (2.44). The larger the values of $r^{(j)}$ are, the higher is the dependency on the *a priori* knowledge, whereas for $r^{(j)} = 1$ there is no influence of φ_ℓ.

2.2.2 HMM Decoder

According to (Rabiner and Juang, 1993, Sec. 6.4) and (Huang et al., 2001, Sec. 8.2), three basic problems of HMMs exist that are adapted as follows to the present ABE framework exploiting phonetic *a priori* knowledge:

1. Evaluation problem – How can the joint production likelihood $\mathrm{p}(\mathbf{x}_0^{L-1}, \varphi_0^{L-1}|\lambda)$ of the observation sequences $\mathbf{x}_0^{L-1} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{L-1})$ and $\varphi_0^{L-1} = (\varphi_0, \varphi_1, \ldots, \varphi_{L-1})$ given model λ be efficiently computed?
2. Decoding problem – How can a state sequence $(s^*)_0^{L-1} = (s_0^*, s_1^*, \ldots, s_{L-1}^*)$ be found that optimally explains the observation sequences $\mathbf{x}_0^{L-1}$ and φ_0^{L-1} given model λ?
3. Learning problem – How can the model parameters $\lambda = \{\pi_i, a_{i,j}, b_j(\mathbf{x}) \mid i, j \in \mathcal{S}\}$ be trained maximizing $\mathrm{p}(\mathbf{x}_0^{L-1}, \varphi_0^{L-1}|\lambda)$?

The learning problem has already been solved by means of an ML parameter estimation via the commonly used EM algorithm in Sec. 2.1.4, according to (Rabiner and Juang,

1993, Sec. 6.4.3). Please note that the state observation likelihoods have been modeled by GMMs that only depend on the feature observations, whereas the phonetic observations are introduced by means of a phoneme class probability matrix defined in Sec. 2.2.1.

Proposals to tackle both remaining problems are provided in the following two subsections. While the evaluation problem can be solved via an FA- or FBA-based *optimal state decoder* (Rabiner and Juang, 1993, Sec. 6.4.1), the answer to the decoding problem is given by a VA-based *optimal state sequence decoder* (Rabiner and Juang, 1993, Sec. 6.4.2).

Optimal State Decoder

The HMM evaluation problem can be tackled via the well-known FBA (Bahl et al., 1974), which is also known as BCJR algorithm due to its inventors Bahl, Cocke, Jelinek, and Raviv. Being divided into a forward and backward recursion, the FBA provides the locally optimal state. An efficient solution can be alternatively obtained by means of the FA requiring only a forward recursion.

On the one hand, the *forward recursion* is defined by means of a forward variable $\alpha_\ell(i) = \mathrm{p}(\mathbf{x}_0^\ell, \varphi_0^\ell, s_\ell = i|\lambda)$ that denotes the joint PDF of the partial observation sequences $\mathbf{x}_0^\ell = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_\ell)$ and $\varphi_0^\ell = (\varphi_0, \varphi_1, \ldots, \varphi_\ell)$ as well as of state $s_\ell = i$ given the model λ. Following (Rabiner and Juang, 1993, Sec. 6.4.1.1), the forward recursion thereby results in

$$\begin{aligned}\alpha_{\ell+1}(j) &= b_j(\mathbf{x}_{\ell+1}, \varphi_{\ell+1}) \sum_{i=0}^{N_\mathcal{S}-1} a_{i,j}\, \alpha_\ell(i) \\ &= b_j(\mathbf{x}_{\ell+1})\, \mathrm{P}(\varphi_{\ell+1}|s_{\ell+1}=j) \sum_{i=0}^{N_\mathcal{S}-1} a_{i,j}\, \alpha_\ell(i) \quad \forall\, j \in \mathcal{S},\ \ell = 0, 1, \ldots, L-2.\end{aligned} \tag{2.47}$$

For initialization, the forward variable is set to $\alpha_0(i) = \pi_i\, b_i(\mathbf{x}_0, \varphi_0) = \pi_i\, b_i(\mathbf{x}_0)\, \mathrm{P}(\varphi_0|s_0 = i)$ $\forall\, i \in \mathcal{S}$. In case of using no phonetic observations, (2.47) can be simplified to

$$\alpha_{\ell+1}(j) = b_j(\mathbf{x}_{\ell+1}) \sum_{i=0}^{N_\mathcal{S}-1} a_{i,j}\, \alpha_\ell(i) \quad \forall\, j \in \mathcal{S},\ \ell = 0, 1, \ldots, L-2, \tag{2.48}$$

with $\alpha_\ell(i) = \mathrm{p}(\mathbf{x}_0^\ell, s_\ell = i|\lambda)$ being a forward variable that is now independent from φ_0^ℓ and thus initialized by $\alpha_0(i) = \pi_i\, b_i(\mathbf{x}_0)$ $\forall\, i \in \mathcal{S}$.

On the other hand, the *backward recursion* is based on a backward variable $\beta_\ell(i) = \mathrm{p}(\mathbf{x}_{\ell+1}^{L-1}, \varphi_{\ell+1}^{L-1}|s_\ell = i, \lambda)$. In contrast to the forward variable, it only involves the sequences of future observations $\mathbf{x}_{\ell+1}^{L-1} = (\mathbf{x}_{\ell+1}, \mathbf{x}_{\ell+2}, \ldots, \mathbf{x}_{L-1})$ and $\varphi_{\ell+1}^{L-1} = (\varphi_{\ell+1}, \varphi_{\ell+2}, \ldots, \varphi_{L-1})$. Furthermore, the backward variable is conditioned by the state $s_\ell = i$ in addition to the model λ and initialized by $\beta_{L-1}(j) = 1$ $\forall\, j \in \mathcal{S}$. Following (Rabiner and Juang, 1993, Sec. 6.4.1.2),

the backward recursion thereby results in

$$\begin{aligned}\beta_\ell(i) &= \sum_{j=0}^{N_S-1} b_j(\mathbf{x}_{\ell+1}, \varphi_{\ell+1})\, a_{i,j}\, \beta_{\ell+1}(j) \\ &= \sum_{j=0}^{N_S-1} b_j(\mathbf{x}_{\ell+1})\, \mathrm{P}(\varphi_{\ell+1}|s_{\ell+1}=j)\, a_{i,j}\, \beta_{\ell+1}(j) \quad \forall\, i \in \mathcal{S},\ \ell = L-2, L-3, \ldots, 0.\end{aligned} \tag{2.49}$$

When using no phonetic observations, the backward variable does not depend on $\varphi_{\ell+1}^{L-1}$ and is therefore defined as $\beta_\ell(i) = \mathrm{p}(\mathbf{x}_{\ell+1}^{L-1}|s_\ell = i, \lambda)$. This has no effect on the initial values $\beta_{L-1}(j) = 1\ \forall\ j \in \mathcal{S}$. The simplified version of (2.49) then yields

$$\beta_\ell(i) = \sum_{j=0}^{N_S-1} b_j(\mathbf{x}_{\ell+1})\, a_{i,j}\, \beta_{\ell+1}(j) \quad \forall\, i \in \mathcal{S},\ \ell = L-2, L-3, \ldots, 0. \tag{2.50}$$

Coming back to the HMM evaluation problem, the required joint production likelihood can be efficiently calculated by definition of the terminal forward variable just via the complete forward recursion (Rabiner and Juang, 1993, Eq. (6.21))

$$\begin{aligned}\mathrm{p}(\mathbf{x}_0^{L-1}, \varphi_0^{L-1}|\lambda) &= \sum_{j=0}^{N_S-1} \mathrm{p}(\mathbf{x}_0^{L-1}, \varphi_0^{L-1}, s_{L-1}=j|\lambda) \\ &= \sum_{j=0}^{N_S-1} \alpha_{L-1}(j).\end{aligned} \tag{2.51}$$

Alternatively, the forward and backward recursions can be used together to compute the joint production likelihood at any frame ℓ (Pfister and Kaufmann, 2008, Eqs. (104)–(105))

$$\begin{aligned}\mathrm{p}(\mathbf{x}_0^{L-1}, \varphi_0^{L-1}|\lambda) &= \sum_{i=0}^{N_S-1} \mathrm{p}(\mathbf{x}_0^{L-1}, \varphi_0^{L-1}, s_\ell=i|\lambda) \\ &= \sum_{i=0}^{N_S-1} \mathrm{p}(\mathbf{x}_{\ell+1}^{L-1}, \varphi_{\ell+1}^{L-1}, \mathbf{x}_0^{\ell}, \varphi_0^{\ell}, s_\ell=i|\lambda) \\ &= \sum_{i=0}^{N_S-1} \mathrm{p}(\mathbf{x}_{\ell+1}^{L-1}, \varphi_{\ell+1}^{L-1}|\mathbf{x}_0^{\ell}, \varphi_0^{\ell}, s_\ell=i, \lambda)\, \mathrm{p}(\mathbf{x}_0^{\ell}, \varphi_0^{\ell}, s_\ell=i|\lambda) \\ &= \sum_{i=0}^{N_S-1} \mathrm{p}(\mathbf{x}_{\ell+1}^{L-1}, \varphi_{\ell+1}^{L-1}|s_\ell=i, \lambda)\, \mathrm{p}(\mathbf{x}_0^{\ell}, \varphi_0^{\ell}, s_\ell=i|\lambda) \\ &= \sum_{i=0}^{N_S-1} \beta_\ell(i)\, \alpha_\ell(i) \quad \forall\ \ell = 0, 1, \ldots, L-1.\end{aligned} \tag{2.52}$$

Having a look on the dependency scheme in Fig. 2.8 and taking into account (S. Walz, 2011, Eq. (2.17)), the simplification $\mathrm{p}(\mathbf{x}_{\ell+1}^{L-1}, \varphi_{\ell+1}^{L-1}|\mathbf{x}_0^{\ell}, \varphi_0^{\ell}, s_\ell=i, \lambda) = \mathrm{p}(\mathbf{x}_{\ell+1}^{L-1}, \varphi_{\ell+1}^{L-1}|s_\ell=i, \lambda)$ is based on the assumption that for the given state $s_\ell=i$ the subsequent observation sequences $\mathbf{x}_{\ell+1}^{L-1}$ and $\varphi_{\ell+1}^{L-1}$ are *conditionally independent* from their preceding observation sequences $\mathbf{x}_0^{\ell}$ and φ_0^{ℓ}, respectively.

The *a posteriori* probabilities $\boldsymbol{\gamma}_\ell = [\gamma_\ell(0), \gamma_\ell(1), \ldots, \gamma_\ell(N_\mathcal{S}-1)]^\mathrm{T}$ mentioned in Fig. 2.7 can be computed either via the entire FBA using (2.47) and (2.49) (Rabiner and Juang, 1993, Eqs. (6.26)–(6.28))

$$\begin{aligned}\gamma_\ell(i) &= \mathrm{P}(s_\ell = i|\mathbf{x}_0^{L-1}, \boldsymbol{\varphi}_0^{L-1}, \lambda) \\ &= \frac{\mathrm{p}(\mathbf{x}_0^{L-1}, \boldsymbol{\varphi}_0^{L-1}, s_\ell = i|\lambda)}{\mathrm{p}(\mathbf{x}_0^{L-1}, \boldsymbol{\varphi}_0^{L-1}|\lambda)} \\ &= \frac{\alpha_\ell(i)\,\beta_\ell(i)}{\sum_{j=0}^{N_\mathcal{S}-1}\alpha_\ell(j)\,\beta_\ell(j)} \quad \forall\, i \in \mathcal{S},\ \ell = 0, 1, \ldots, L-1,\end{aligned} \tag{2.53}$$

or by means of the FA just using (2.47) (Bauer et al., 2014a, Eqs. (1)–(2))

$$\begin{aligned}\gamma_\ell(i) &= \mathrm{P}(s_\ell = i|\mathbf{x}_0^{\ell}, \boldsymbol{\varphi}_0^{\ell}, \lambda) \\ &= \frac{\mathrm{p}(\mathbf{x}_0^{\ell}, \boldsymbol{\varphi}_0^{\ell}, s_\ell = i|\lambda)}{\mathrm{p}(\mathbf{x}_0^{\ell}, \boldsymbol{\varphi}_0^{\ell}|\lambda)} \\ &= \frac{\alpha_\ell(i)}{\sum_{j=0}^{N_\mathcal{S}-1}\alpha_\ell(j)} \quad \forall\, i \in \mathcal{S},\ \ell = 0, 1, \ldots, L-1.\end{aligned} \tag{2.54}$$

When making no use of phonetic observations, the calculation of the *a posteriori* probabilities is simplified by employing the forward/backward recursions in (2.48) and (2.50). Thus, the FBA-based equation (2.53) turns into

$$\begin{aligned}\gamma_\ell(i) &= \mathrm{P}(s_\ell = i|\mathbf{x}_0^{L-1}, \lambda) \\ &= \frac{\mathrm{p}(\mathbf{x}_0^{L-1}, s_\ell = i|\lambda)}{\mathrm{p}(\mathbf{x}_0^{L-1}|\lambda)} \\ &= \frac{\alpha_\ell(i)\,\beta_\ell(i)}{\sum_{j=0}^{N_\mathcal{S}-1}\alpha_\ell(j)\,\beta_\ell(j)} \quad \forall\, i \in \mathcal{S},\ \ell = 0, 1, \ldots, L-1,\end{aligned} \tag{2.55}$$

while the FA-based equation (2.54) results in

$$\begin{aligned}\gamma_\ell(i) &= \mathrm{P}(s_\ell = i|\mathbf{x}_0^{\ell}, \lambda) \\ &= \frac{\mathrm{p}(\mathbf{x}_0^{\ell}, s_\ell = i|\lambda)}{\mathrm{p}(\mathbf{x}_0^{\ell}|\lambda)} \\ &= \frac{\alpha_\ell(i)}{\sum_{j=0}^{N_\mathcal{S}-1}\alpha_\ell(j)} \quad \forall\, i \in \mathcal{S},\ \ell = 0, 1, \ldots, L-1.\end{aligned} \tag{2.56}$$

Please note that the stochastic constraint $\sum_{i=0}^{N_\mathcal{S}-1}\gamma_\ell(i) = 1$ is met for (2.53)–(2.56).

The aforementioned decoding problem can be already solved 'locally' by the FA- or FBA-based optimal state decoder. Thus, the individually most likely *state* at frame ℓ corresponds to the maximum *a posteriori* (MAP) probability (Rabiner and Juang, 1993, Eq. (6.29))

$$s_\ell^* = \underset{i\in\{0,\ldots,N_\mathcal{S}-1\}}{\arg\max}\ \gamma_\ell(i) \quad \forall\, \ell = 0, 1, \ldots, L-1. \tag{2.57}$$

However, the *sequence* of 'globally' optimum states is thereby not obtained (Rabiner and Juang, 1993, Sec. 6.4.2).

Optimal State Sequence Decoder

A popular way to solve the decoding problem is the VA (Rabiner and Juang, 1993, Sec. 6.4.2.1), which is named after its inventor Viterbi (1967). In contrast to (2.57), it derives the optimal state *sequence* $(s^*)_0^{L-1} = (s_0^*, s_1^*, \ldots, s_{L-1}^*)$ from all possible state sequences $s_0^{L-1} = (s_0, s_1, \ldots, s_{L-1})$ by means of the observation sequences $\mathbf{x}_0^{L-1}$ and φ_0^{L-1} as well as the model λ (Rabiner and Juang, 1993, Sec. 6.4.2):

$$\begin{aligned}(s^*)_0^{L-1} &= \arg\max_{s_0^{L-1}} \mathrm{P}(s_0^{L-1}|\mathbf{x}_0^{L-1}, \varphi_0^{L-1}, \lambda) \\ &= \arg\max_{s_0^{L-1}} \mathrm{P}(s_0^{L-1}, \mathbf{x}_0^{L-1}, \varphi_0^{L-1}|\lambda).\end{aligned} \tag{2.58}$$

When making no use of phonetic observations, (2.58) turns into

$$(s^*)_0^{L-1} = \arg\max_{s_0^{L-1}} \mathrm{P}(s_0^{L-1}, \mathbf{x}_0^{L-1}|\lambda). \tag{2.59}$$

The Viterbi score along the best of the partial state sequences $s_0^{\ell-1} = (s_0, s_1, \ldots, s_{\ell-1})$ ending up in state $s_\ell = i$ for the observation sequences $\mathbf{x}_0^\ell$ and φ_0^ℓ given the model λ is defined as (Rabiner and Juang, 1993, Eq. (6.30))

$$\delta_\ell(i) = \max_{s_0^{\ell-1}} \mathrm{P}(s_0^{\ell-1}, s_\ell = i, \mathbf{x}_0^\ell, \varphi_0^\ell|\lambda). \tag{2.60}$$

All Viterbi scores need to be recursively calculated similar to the forward recursion (2.47), with the sum being replaced by a maximization (Rabiner and Juang, 1993, Eq. (6.33a))

$$\begin{aligned}\delta_{\ell+1}(j) &= b_j(\mathbf{x}_{\ell+1}, \varphi_{\ell+1}) \max_{i\in\{0,\ldots,N_\mathcal{S}-1\}} \left[a_{i,j}\, \delta_\ell(i)\right] \\ &= b_j(\mathbf{x}_{\ell+1})\, \mathrm{P}(\varphi_{\ell+1}|s_{\ell+1} = j) \max_{i\in\{0,\ldots,N_\mathcal{S}-1\}} \left[a_{i,j}\, \delta_\ell(i)\right] \quad \forall\, j \in \mathcal{S},\ \ell = 0, 1, \ldots, L-2.\end{aligned} \tag{2.61}$$

The intial Viterbi score is also initialized by $\delta_0(i) = \pi_i\, b_i(\mathbf{x}_0, \varphi_0) = \pi_i\, b_i(\mathbf{x}_0)\, \mathrm{P}(\varphi_0|s_0 = i)$ $\forall\ i \in \mathcal{S}$. Simultaneously, the arguments of the maximization in (2.61) are stored by a backtracking pointer (Rabiner and Juang, 1993, Eq. (6.33b))

$$\psi_{\ell+1}(j) = \arg\max_{i\in\{0,\ldots,N_\mathcal{S}-1\}} \left[a_{i,j}\, \delta_\ell(i)\right] \quad \forall\, j \in \mathcal{S},\ \ell = 0, 1, \ldots, L-2. \tag{2.62}$$

The Viterbi recursion is terminated by setting the optimum state of the last frame to $s_{L-1}^* = \arg\max_{j\in\{0,\ldots,N_\mathcal{S}-1\}} \delta_{L-1}(j)$ (Rabiner and Juang, 1993, Eq. (6.34b)). Subsequently, the optimal state sequence is decoded via a recursive backtracking procedure (Rabiner and Juang, 1993, Eq. (6.35))

$$s_\ell^* = \psi_{\ell+1}(s_{\ell+1}^*) \quad \forall\, \ell = L-2, L-3, \ldots, 0. \tag{2.63}$$

In case of using no phonetic observations, the definition of the Viterbi score (2.60) yields

$$\delta_\ell(i) = \max_{s_0^{\ell-1}} \mathrm{P}(s_0^{\ell-1}, s_\ell = i, \mathbf{x}_0^\ell|\lambda) \tag{2.64}$$

and the recursive computation of the Viterbi scores in (2.61) is simplified to

$$\delta_{\ell+1}(j) = b_j(\mathbf{x}_{\ell+1}) \max_{i\in\{0,\ldots,N_\mathcal{S}-1\}} \left[a_{i,j}\,\delta_\ell(i)\right] \quad \forall\, j \in \mathcal{S},\ \ell = 0, 1, \ldots, L-2, \tag{2.65}$$

with $\delta_0(i) = \pi_i\, b_i(\mathbf{x}_0)\ \forall\ i \in \mathcal{S}$ denoting the initial Viterbi score. Based on the adapted Viterbi scores, the derivation of the backtracking pointers in (2.62) as well as the decoding of the optimal state sequence in (2.63) remain unchanged.

Note that when using the VA, there is no practically feasible exact way to obtain $\gamma_\ell(i)$.

2.2.3 Estimation of UB Cepstral Envelope

As depicted in Fig. 2.7, the UB cepstral envelope $\hat{\mathbf{c}}_\ell$ is estimated by means of the pre-trained CB $\mathcal{C} = \left\{\mathbf{c}^{(i)} \mid i \in \mathcal{S}\right\}$, with $\mathbf{c}^{(i)}\ \forall\ i = 0, 1, \ldots, N_\mathcal{S} - 1$ denoting the CB entries. On the one hand, the *a posteriori* probabilities $\boldsymbol{\gamma}_\ell = [\gamma_\ell(0), \gamma_\ell(1), \ldots, \gamma_\ell(N_\mathcal{S}-1)]^\mathrm{T}$ derived from the FA or FBA can thereby be used for maximum *a posteriori* (MAP) or minimum mean square error (MMSE) estimation (Jax, 2002, Sec. 6.4.3–6.4.4). On the other hand, a Viterbi path estimation[10] can be utilized by employing the state s_ℓ^*, which belongs to the optimal state sequence $(s^*)_0^{L-1} = (s_0^*, s_1^*, \ldots, s_\ell^*, \ldots, s_{L-1}^*)$ decoded by the VA. In the following, all three estimators will be briefly introduced.

MAP Estimation

The MAP estimation rule selects a CB entry of $\mathcal{C}$ corresponding to the highest *a posteriori* probability provided by the FA or FBA. Thus, the 'locally' optimum state s_ℓ^* obtained in (2.57) just acts as an index to the CB entries (Jax, 2002, Sec. 6.4.3)

$$\hat{\mathbf{c}}_\ell = \mathbf{c}^{(s_\ell^*)} = E\left\{\mathbf{c}_\ell \mid s_\ell^*\right\}. \tag{2.66}$$

However, potential representations of the UB cepstral envelope are therefore limited to the discrete CB entries.

MMSE Estimation

To allow for a continuous UB cepstral envelope estimation, the *a posteriori* probabilities, which are computed by means of the FA in (2.54)/(2.56) or the FBA in (2.53)/(2.55), can

[10] Please refer to Yagli et al. (2013), when restricting the temporal look-ahead for Viterbi path estimation.

serve as weights of the CB entries (Jax, 2002, Eq. (6.36))

$$\hat{\mathbf{c}}_\ell = \sum_{i=0}^{N_\mathcal{S}-1} \mathbf{c}^{(i)} \gamma_\ell(i). \tag{2.67}$$

This leads to an MMSE criterion according to (Jax, 2002, Sec. 6.4.4). Theoretically, an infinite number of cepstral estimates can thereby be realized.

Viterbi Path Estimation

Similar to (2.66), the Viterbi path estimator selects a discrete CB entry of $\mathcal{C}$, which corresponds this time to the optimal state sequence decoded by the VA in (2.63)

$$\hat{\mathbf{c}}_\ell = \mathbf{c}^{(s_\ell^*)} = E\left\{ \mathbf{c}_\ell \,\middle|\, (s^*)_0^{L-1} = (s_0^*, s_1^*, \ldots, s_\ell^*, \ldots, s_{L-1}^*) \right\}. \tag{2.68}$$

Thus, the UB cepstral envelope is represented by the 'globally' optimum CB entry compared to the more 'locally' oriented MAP and MMSE estimations (2.66)–(2.67).

2.2.4 Conversion to UB Spectral Envelope

As depicted in Fig. 2.7, the estimated UB cepstral envelope $\hat{\mathbf{c}}_\ell$ needs to be converted into an UB short-term power spectrum $\hat{\mathbf{\Phi}}_{\mathrm{UB},\ell}$. In contrast to (Jax, 2002, Sec. 6.1.2), this spectral conversion is conducted under a moderate cepstral interframe smoothing constraint (Bauer and Fingscheidt, 2009a,b). Transients as well as discontinuities, which may potentially arise from the time-variant WB LP analysis and synthesis filters provoking undesired artifacts (Schnell and Lacroix, 2008; Välimäki, 1995; Välimäki and Laakso, 1998), will be thereby reduced beforehand. These switching effects are empirically detected, if the LPCC-based log-spectral distance (LSD) (Jax, 2002, Eq. (4.15)) between the estimated UB cepstral envelope vectors in two successive frames

$$\mathit{LSD}_{\mathrm{ceps},\ell} = \frac{10\sqrt{2}}{\ln 10} \left\| \hat{\mathbf{c}}_\ell - \hat{\mathbf{c}}_{\ell-1} \right\|, \quad 0 < \ell \leqslant L-1, \tag{2.69}$$

with $\|\cdot\|$ denoting the Euclidean vector norm, exceeds a predefined threshold of 30 dB. In case of a detection $\mathit{LSD}_{\mathrm{ceps},\ell} > 30$ dB, the estimated UB cepstral envelope vector of the current frame is recursively smoothed by its predecessor following the redefinition

$$\hat{\mathbf{c}}_\ell := \frac{\hat{\mathbf{c}}_\ell + \hat{\mathbf{c}}_{\ell-1}}{2} \tag{2.70}$$

until $\mathit{LSD}_{\mathrm{ceps},\ell} < 20$ dB holds. From experience, the cepstral interframe smoothing takes place only during hard switching effects and leads to a moderate reduction of artifacts without interfering too much.

The subsequent spectral conversion is done according to (Jax, 2002, Eqs. (6.2)–(6.3))

$$\hat{\Phi}_{\mathrm{UB},\ell}(\tilde{k}) = \exp\left(2\Re\left\{\frac{\hat{c}_\ell(0)}{\sqrt{2}} + \sum_{\tilde{n}=1}^{K_{\mathrm{UB}}-1} \hat{c}_\ell(\tilde{n}) \cdot e^{-j2\pi\frac{\tilde{n}\tilde{k}}{K_{\mathrm{UB}}}}\right\}\right), \quad \tilde{k} = 0, 1, \ldots, K_{\mathrm{UB}} - 1, \quad (2.71)$$

with $\Re\{\cdot\}$ denoting the real part of its complex argument. When dividing the zeroth LPCC by $\sqrt{2}$ and padding $\hat{c}_\ell(\tilde{n})\ \forall\ \tilde{n} = N_{\mathrm{LP}^{(\mathrm{UB})}}+1, N_{\mathrm{LP}^{(\mathrm{UB})}}+2, \ldots, K_{\mathrm{UB}}-1$ with zeros, the argument of $\Re\{\cdot\}$ can be realized by a short-term DFT and efficiently computed via an FFT implementation (Oppenheim and Schafer, 1989, Sec. 8.1 and Sec. 9.1). Considering (2.12), $\hat{\mathbf{\Phi}}_{\mathrm{UB},\ell}$ represents the estimated spectral envelope of the critically downsampled UB short-term power spectrum with a normalized energy according to (2.17).

2.2.5 WB Spectral Assembly

Before being able to assemble the estimated WB short-term power spectrum $\hat{\mathbf{\Phi}}_{\mathrm{WB},\ell}$, the NB short-term power spectrum $\mathbf{\Phi}_{\mathrm{NB},\ell}$ is consistently calculated according to (2.9) and (2.11) based on the NB input speech samples $s_{\mathrm{NB}}(n')$, as depicted in Fig. 2.7. For a correct frequency bin allocation from $\Phi_{\mathrm{NB},\ell}(k')$ to $\hat{\Phi}_{\mathrm{WB},\ell}(k)$, a mapping function $\mathrm{k}_{\mathrm{NB}}(\cdot)$ is used. It maps the domain $\tilde{\mathcal{K}}_{\mathrm{NB}} = \{0, 1, \ldots, N'_{\mathrm{w}} - 1\}$ of frequency bins k' to the range $\mathcal{K}_{\mathrm{NB}} = \left\{0, 1, \ldots, \frac{N'_{\mathrm{w}}}{2} - 1, N_{\mathrm{w}} - \frac{N'_{\mathrm{w}}}{2}, \ldots, N_{\mathrm{w}} - 1\right\}$ of frequency bins $k = \mathrm{k}_{\mathrm{NB}}(k')$:

$$\mathrm{k}_{\mathrm{NB}} \ :\ \tilde{\mathcal{K}}_{\mathrm{NB}} \mapsto \mathcal{K}_{\mathrm{NB}} \ :\ k' \mapsto \mathrm{k}_{\mathrm{NB}}(k') = k. \quad (2.72)$$

Furthermore, a LB LP analysis, as shown in Fig. 2.3, is performed by computing the first $N_{\mathrm{LP}^{(\mathrm{LB})}}+1$ ACF coefficients $\phi_{\mathrm{LB},\ell}(\tilde{n}')\ \forall\ \tilde{n}' \in \{0, 1, \ldots, N_{\mathrm{LP}^{(\mathrm{LB})}}\}$ according to (2.13) and applying the well-known Levinson-Durbin recursion (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978). As a result, the LB prediction gain factor $\sigma_{\mathrm{LB},\ell}$ is obtained to compensate for the energy normalization of the estimated UB spectral envelope $\hat{\mathbf{\Phi}}_{\mathrm{UB},\ell}$.

The spectral assembly can be thereby conducted following (Jax, 2002, Eq. (6.5))

$$\hat{\Phi}_{\mathrm{WB},\ell}(k) = \begin{cases} \hat{\Phi}_{\mathrm{UB},\ell}\left(\mathrm{k}_{\mathrm{UB}}^{-1}(k)\right) \cdot \sigma^2_{\mathrm{LB},\ell}, & \text{if } k \in \mathcal{K}_{\mathrm{UB}}, \\ \Phi_{\mathrm{NB},\ell}\left(\mathrm{k}_{\mathrm{NB}}^{-1}(k)\right), & \text{if } k \in \mathcal{K}_{\mathrm{NB}} \setminus \mathcal{K}_{\mathrm{UB}}, \end{cases} \quad (2.73)$$

with $\mathrm{k}_{\mathrm{UB}}^{-1}(\cdot)$ and $\mathrm{k}_{\mathrm{NB}}^{-1}(\cdot)$ denoting the inverse[11] of the mapping function $\mathrm{k}_{\mathrm{UB}}(\cdot)$ in (2.14) and $\mathrm{k}_{\mathrm{NB}}(\cdot)$ in (2.72), respectively. Accordingly, the union $\mathcal{K}_{\mathrm{UB}} \bigcup \mathcal{K}_{\mathrm{NB}}$ is equal to $\mathcal{K}_{\mathrm{WB}} = \{0, 1, \ldots, N_{\mathrm{w}} - 1\}$ comprising the frequency bins of $\hat{\mathbf{\Phi}}_{\mathrm{WB},\ell}$. The intersection $\mathcal{K}_{\mathrm{UB}} \bigcap \mathcal{K}_{\mathrm{NB}}$, however, is only empty, if the frequency bands are specified by a cut-off frequency of $f_{\mathrm{c}} = f'_{\mathrm{s}}/2$ in Sec. 2.1.2. Hence, (2.73) generally takes into account the complement $\mathcal{K}_{\mathrm{NB}} \setminus \mathcal{K}_{\mathrm{UB}}$, which only comprises those frequency bins that are dedicated to the LB short-term power spectrum.

[11] Please note that the one-to-one mapping functions $\mathrm{k}_{\mathrm{UB}}(\cdot)$ and $\mathrm{k}_{\mathrm{NB}}(\cdot)$ are bijective and thus invertible.

Focusing on (2.71), the normalization of the zeroth LPCC in (2.17) is compensated for due to the multiplication of $\hat{\tilde{\mathbf{\Phi}}}_{\mathrm{UB},\ell}$ by $\sigma^2_{\mathrm{LB},\ell}$ in (2.73) considering the auxiliary calculation

$$\exp\left(2\Re\left\{\frac{\hat{c}_\ell(0)}{\sqrt{2}}\right\}\right) = \exp\left(\sqrt{2}\hat{c}_\ell(0)\right) = \exp\left(\ln\left(\frac{\hat{\sigma}^2_{\mathrm{UB},\ell}}{\sigma^2_{\mathrm{LB},\ell}}\right)\right) = \frac{\hat{\sigma}^2_{\mathrm{UB},\ell}}{\sigma^2_{\mathrm{LB},\ell}}, \tag{2.74}$$

with $\hat{\sigma}_{\mathrm{UB},\ell}$ denoting the estimated UB prediction gain factor.

2.2.6 Conversion to WB LPC Coefficients

The spectrally assembled estimated WB short-term power spectrum $\hat{\tilde{\mathbf{\Phi}}}_{\mathrm{WB},\ell}$ finally needs to be converted into the desired WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$. A WB LP analysis is therefore required as depicted in Fig. 2.7. It employs a short-term inverse DFT similar to (2.12)–(2.13), but this time based on the WB instead of a sub-band short-term power spectrum (Oppenheim and Schafer, 1989, Sec. 8.1)

$$\hat{\phi}_{\mathrm{WB},\ell}(n) = \frac{1}{N_{\mathrm{w}}} \sum_{k\in\mathcal{K}_{\mathrm{WB}}} \hat{\Phi}_{\mathrm{WB},\ell}(k)\cdot e^{j2\pi\frac{nk}{N_{\mathrm{w}}}}, \quad n = 0,1,\ldots,N_{\mathrm{LP^{(WB)}}}, \tag{2.75}$$

with the WB LP order being set to $N_{\mathrm{LP^{(WB)}}} = 16$, i.e., $N_{\mathrm{LP^{(WB)}}} < N_{\mathrm{w}} - 1$. The truncated estimated WB ACF $\hat{\boldsymbol{\phi}}_{\mathrm{WB},\ell}$ is then fed into the well-known Levinson-Durbin recursion yielding the estimated WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell} = [\hat{a}_{\mathrm{WB},\ell}(1), \hat{a}_{\mathrm{WB},\ell}(2), \ldots, \hat{a}_{\mathrm{WB},\ell}(N_{\mathrm{LP^{(WB)}}})]^{\mathrm{T}}$ (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978). Additionally, the corresponding reflection coefficients $\hat{\mathbf{r}}_{\mathrm{WB},\ell} = [\hat{r}_{\mathrm{WB},\ell}(1), \hat{r}_{\mathrm{WB},\ell}(2), \ldots, \hat{r}_{\mathrm{WB},\ell}(N_{\mathrm{LP^{(WB)}}})]^{\mathrm{T}}$ are obtained as a by-product of the Levinson-Durbin recursion. They are used to verify the stability[12] of the autoregressive WB LP synthesis filter as follows (Vary and Martin, 2006, Sec. 6.3.1.3)

$$|\hat{r}_{\mathrm{WB},\ell}(n)| < 1 \quad \forall\, n = 1,2,\ldots,N_{\mathrm{LP^{(WB)}}}. \tag{2.76}$$

As the frame-wise vocal tract estimation following the upper signal paths in Fig. 2.7 is hereby finished, the sample-wise main ABE processing based on the lower signal path is detailed now in the remaining part of Sec. 2.2.

2.2.7 Interpolation

Before employing the estimated WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$, the NB input speech samples $s_{\mathrm{NB}}(n')$ first need to be interpolated from $f'_{\mathrm{s}} = 8$ kHz to $f_{\mathrm{s}} = 16$ kHz, as depicted in Fig. 2.7. The interpolation implies an upsampling of factor two and a subsequent lowpass filtering (Oppenheim and Schafer, 1989; Rabiner and Schafer, 1978; Proakis and Manolakis, 2007).

[12] Please note that filter instabilities have never been encountered in our simulations with 64 bit precision.

In a first step, the upsampling is realized by inserting after each sample a zero

$$s_{\mathrm{NB}}(n) = \begin{cases} 2 \cdot s_{\mathrm{NB}}(n'), & \text{if } n' = \frac{n}{2} \in \mathbb{Z}, \\ 0, & \text{else}, \end{cases} \tag{2.77}$$

with $s_{\mathrm{NB}}(n)$ denoting the upsampled version of the NB input speech $s_{\mathrm{NB}}(n')$. To maintain the signal energy after interpolation, the non-zero samples need to be multiplied by two.

After that, an anti-aliasing lowpass filter is applied to preserve the unmodified LB spectrum, while removing the undesired spectral components from the UB frequency range. The cut-off frequency of the lowpass filter is therefore adapted to f_{c} specifying the frequency bands in Sec. 2.1.2. In this work, a linear-phase FIR lowpass filter of order N_{LB} is employed. Its impulse response is represented by the filter coefficients $\mathbf{b}_{\mathrm{LB}} = [b_{\mathrm{LB}}(0), b_{\mathrm{LB}}(1), \ldots, b_{\mathrm{LB}}(N_{\mathrm{LB}})]^{\mathrm{T}}$. An FIR lowpass filtering of $s_{\mathrm{NB}}(n)$ thereby yields the interpolated LB speech samples

$$s_{\mathrm{LB}}(n) = \sum_{\nu=0}^{N_{\mathrm{LB}}} b_{\mathrm{LB}}(\nu) \cdot s_{\mathrm{NB}}(n-\nu), \tag{2.78}$$

with $s_{\mathrm{NB}}(n) = 0 \; \forall \; n < 0$, i.e., the delay units (memory) of the filter are initialized by zero.

The linear phase of the FIR lowpass filter allows for a constant group delay of $\tau_{\mathrm{LB}} = \frac{N_{\mathrm{LB}}}{2}$ samples. Thus, the algorithmic delay contributions in the different signal paths of Fig. 2.7 can be exactly compensated for to ensure time alignment on the basis of frames and samples, respectively. This plays an important role, particularly for the online ABE application in Sec. 4.

2.2.8 LP Analysis and Synthesis Filtering

Following the well-known source-filter model of human speech production (Flanagan, 1972), a speech signal can be divided into an *excitation* at the glottis and a *vocal tract filter* neglecting the nasal tract:

- The excitation is generated by pressuring air via the lungs through the glottis. In case of unvoiced sounds, the vocal cords surrounding the glottis are opened, as while breathing. The excitation is therefore noisy and spectrally flat. In contrast, the vocal cords are almost closed and put into vibration for voiced sounds. The vibration rate (periodicity) is thereby specified by the fundamental frequency F_0, which is colloquially called pitch. It depends on the speaker and the emotional manner of speaking. The spectrum of the excitation therefore reveals different equidistant comb structures with maxima at F_0 and integer multiples of it (i.e., harmonics). While male and female voices reveal fundamental frequency ranges of about $50 \ldots 250$ Hz and $120 \ldots 500$ Hz, respectively, the highest F_0 of up to 600 Hz is reached by children (Vary and Martin, 2006, Sec. 2.2).

- Due to its resonance characteristics, the vocal tract shapes phoneme-specific spectral envelopes. It can physically be approximated from the glottis to the lips of the mouth by lossless acoustic tubes of equal length and different diameters (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978). The number of tubes thereby defines the order of the LP model, i.e., $N_{\mathrm{LP(WB)}}$ in case of the WB LP analysis in Sec. 2.2.6. Based on the estimated WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$, the vocal tract is modeled by an autoregressive IIR LP synthesis filter $(1-\hat{A}(z))^{-1}$, with $\hat{A}(z) = \sum_{\nu=1}^{N_{\mathrm{LP(WB)}}} \hat{a}_{\mathrm{WB},\ell}(\nu) \cdot z^{-\nu}$ denoting the z-transform of the predictor. Furthermore, the same predictor coefficients model the *inverse* vocal tract, however, in terms of a moving-average FIR LP analysis filter $1-\hat{A}(z)$.

As depicted in Fig. 2.7, the WB LP analysis filter $1–\hat{A}(z)$ is first applied to the interpolated LB speech samples $s_{\mathrm{LB}}(n)$ resulting in an interpolated LB residual signal

$$e_{\mathrm{LB}}(n) = s_{\mathrm{LB}}(n) - \sum_{\nu=1}^{N_{\mathrm{LP(WB)}}} \hat{a}_{\mathrm{WB},\ell}(\nu) \cdot s_{\mathrm{LB}}(n-\nu), \tag{2.79}$$

which represents a band-limited excitation of the vocal tract. To cover the whole WB frequency range, $e_{\mathrm{LB}}(n)$ still needs to be extended. There are several ways to accomplish such a residual signal extension that are briefly mentioned in the next section. The WB LP synthesis filter $(1-\hat{A}(z))^{-1}$ finally synthesizes the extended residual signal $\hat{e}_{\mathrm{WB}}(n)$ to obtain the estimated WB speech samples

$$\hat{s}_{\mathrm{WB}}(n) = \hat{e}_{\mathrm{WB}}(n) + \sum_{\nu=1}^{N_{\mathrm{LP(WB)}}} \hat{a}_{\mathrm{WB},\ell}(\nu) \cdot \hat{s}_{\mathrm{WB}}(n-\nu). \tag{2.80}$$

The most important property of this *serial* ABE structure is that the WB LP analysis and synthesis filters operate at the same sampling rate of $f_{\mathrm{s}} = 16\,\mathrm{kHz}$, as discussed in (Jax, 2002, Sec. 2.4). Thus, they are exactly inverse to each other and the LB spectrum can pass through them transparently.

Due to the frame-wise estimation of WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$, they are switched for the sample-wise processing in (2.79) and (2.80) every frame $\ell = 0, 1, \ldots, L-1$ at sample index $n = \ell \cdot N_{\mathrm{s}}$. The implementation of such a *time-variant* filter is crucial, since switching effects – also known as filter ringing – may cause transients or discontinuities and provoke annoying artifacts (Välimäki, 1995; Välimäki and Laakso, 1998; Schnell and Lacroix, 2008). As discovered in (P. Bauer, 2007; Bauer and Fingscheidt, 2008a), these problems particularly arise when using a *transposed* filter structure (Oppenheim and Schafer, 1989, Sec. 6.4). Due to the transposition of the multipliers and delay units, the accumulators contain unexpected terms involving new as well as old predictor coefficients during the transition period of a filter coefficient switch. In contrast, these mixed coefficient terms do not appear in a filter implementation without transposition. It is therefore highly recommended to employ a *non-transposed* filter structure, such as a direct form I (Oppenheim and Schafer, 1989, Fig. 6.10).

However, this does not completely prevent switching effects. To further counteract such problems in advance, a moderate cepstral interframe smoothing has been introduced in Sec. 2.2.4. Additionally, a further smoothing strategy based on the LPC reflection coefficients will be presented in Sec. 4.2.4.

2.2.9 Residual Signal Extension

Between the WB LP analysis and synthesis filtering blocks in Fig. 2.7, the extension of the LB residual signal $e_{\mathrm{LB}}(n)$ takes place. The extended residual signal $\hat{e}_{\mathrm{WB}}(n)$ serves as a WB excitation of the vocal tract. It defines the spectral fine structure of the finally synthesized speech signal. Due to the required transparency property of the serial ABE approach mentioned in the section before, the residual signal extension needs to leave the LB spectrum unchanged. Furthermore, the characteristics of the original excitation for voiced and unvoiced sounds shall be approximated as close as possible. According to (Jax, 2002, Sec. 3), there are several concepts for residual signal extension, which demand more or less complexity to meet these requirements.

Nonlinear Distortions

As a first option, nonlinear distortions can be applied to the LB residual signal, e.g., by using power, saturation or rectification functions (Jax, 2002, Sec. 3.2). This approach follows the probably first ABE proposal at all that has been made by Schmidt (1933) and employs a nonlinear processing (Vary and Martin, 2006, Sec. 10.3.1). However, the nonlinearly distorted residual signals further need to be sophistically post-processed by spectral whitening, gain adaptation, and band-stop filtering (Jax, 2002, Fig. 3.3).

Generation of Noise and Pitch Harmonics

Another method is an explicit generation of noise with a subsequent band-stop filter and gain adaptation (Jax, 2002, Sec. 3.1). It is widely used in speech coding, e.g., by the WB AMR speech codec for frequency components above 6.4 kHz (3GPP TS 26.190, 2001). However, if the missing frequency band is wider, the finally synthesized speech signal sounds rather noisy, particularly during voiced sounds. When using a voiced/unvoiced classifier, this method can be restricted to unvoiced sounds and replaced for voiced sounds with a sinusoidal generation of pitch harmonics (Jax, 2002, Sec. 3.1), based on the principle of harmonic modeling (Carl, 1994, Sec. 4.4.1). However, an harmonic extension requires a precise pitch estimation. Otherwise, the finally synthesized speech signal gives the auditive impression of an additional, simultaneous speaker with a similar pitch. A noise-robust estimation of the

pitch is still challenging (Chan and Hui, 1997; Shahnaz et al., 2007; Thomas et al., 2010; Pulakka et al., 2012d). This fact limits also the use of other pitch-adaptive methods, such as (Jax, 2002, Sec. 3.3.3).

Pitch Doubling

Alternatively, an harmonic extension of the residual signal without the need for estimating the pitch is accomplished by pitch doubling, as proposed in (Jax, 2002, Sec. 3.4). After the pitch has been doubled by downsampling and subsequent time stretching, a highpass filter is required to remove the time-stretched spectral components from the LB spectrum of the extended residual signal (Jax, 2002, Fig. 3.6). However, an auditive impression of a further speaker in the background with a doubled pitch sometimes arises in the finally synthesized speech signal.

Spectral Duplication

Spectral duplication was first proposed by Makhoul and Berouti (1979) and represents a simple time-domain modulation (Jax, 2002, Sec. 3.3). It provides an efficient extension of the LB residual signal by spectrally shifting the LB spectrum depending on the normalized angular modulation frequency Ω_{M}. This assumes the LB residual signal to be spectrally flat, which at least holds for unvoiced sounds. However, a time-domain modulation is also feasible for voiced sounds, based on the assumption that the human ear is rather insensitive to deviations from the original spectral fine structure towards higher frequencies (Jax, 2002, Sec. 3.5). These deviations may take into account spectral gaps as well as variations of the harmonic structure. Thus, the estimated spectral envelope turns out to be more relevant than the extended residual signal in terms of subjective speech quality for high-band ABE (Jax, 2002, Sec. 3.5). The residual signal extension in this work is therefore restricted to the use of an efficient time-domain modulation (Carl, 1994, Fig. 4.33).

As depicted in Fig. 2.7, the extended residual signal $\hat{e}_{\mathrm{WB}}(n)$ results from a sample-wise superposition of the LB residual signal $e_{\mathrm{LB}}(n)$ and an estimated UB residual signal $\hat{e}_{\mathrm{UB}}(n)$:

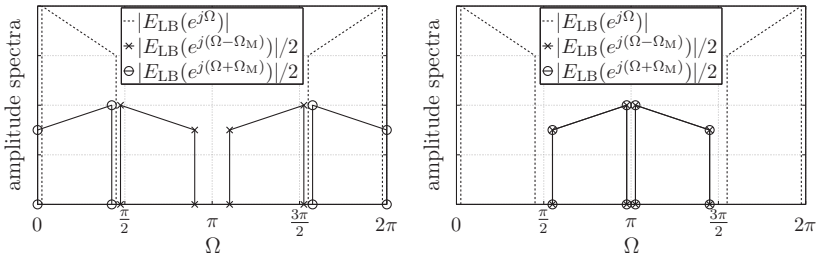
$$\hat{e}_{\mathrm{WB}}(n) = e_{\mathrm{LB}}(n) + \hat{e}_{\mathrm{UB}}(n). \tag{2.81}$$

The latter is obtained by modulating $e_{\mathrm{LB}}(n)$ via a real-valued cosine function (Jax, 2002, Eq. (3.2))

$$\hat{e}_{\mathrm{UB}}(n) = e_{\mathrm{LB}}(n) \cdot g_{\mathrm{M}} \cos(\Omega_{\mathrm{M}} n) \cdot g_{\mathrm{UB}}. \tag{2.82}$$

This time-domain modulation corresponds to a spectral translation (ST) (Jax, 2002, Eq. (3.3))

$$\hat{E}_{\mathrm{UB}}\big(e^{j\Omega}\big) = \frac{g_{\mathrm{M}}}{2}\left[E_{\mathrm{LB}}\big(e^{j(\Omega-\Omega_{\mathrm{M}})}\big) + E_{\mathrm{LB}}\big(e^{j(\Omega+\Omega_{\mathrm{M}})}\big)\right] \cdot g_{\mathrm{UB}}, \tag{2.83}$$



(a) ST: $\Omega_{\mathrm{M}} = \Omega_{\mathrm{c}}$ (b) SF: $\Omega_{\mathrm{M}} = \pi$

Figure 2.9: Schematically illustrated time-domain modulations ST and SF for residual signal extension: The LB amplitude spectrum $|E_{\mathrm{LB}}(e^{j\Omega})|$ given $\Omega_{\mathrm{c}} = 0.45\pi$ is dashed, whereas its shifted halved replica $|E_{\mathrm{LB}}(e^{j(\Omega\pm\Omega_{\mathrm{M}})})|/2$ are marked with asterisks and circles, respectively.

where two copies of the halved LB spectrum $E_{\mathrm{LB}}(e^{j\Omega})/2$ are shifted by $\pm\Omega_{\mathrm{M}}$. Depending on Ω_{M}, some spectral components may be shifted again into the LB frequency range. In these cases, they need to be removed from $\hat{e}_{\mathrm{UB}}(n)$ by means of the anti-aliasing highpass filter shown in Fig. 2.7. A low-delay IIR filter design is thereby recommended to allow for synchronicity between the WB LP analysis and synthesis filtering. Furthermore, an energy adaptation of the UB spectrum is required that is taken into account by the modulation gain g_{M}. In contrast to (Jax, 2002, Fig. 3.4), the UB residual signal is further weighted by an additional static attenuation factor $g_{\mathrm{UB}} = 10^{\frac{g_{\mathrm{UB,dB}}}{20}}$, with the logarithmic weight $g_{\mathrm{UB,dB}}$ representing a negative value in dB. Thus, the aggressiveness of the ABE can be controlled. Another adaptive attenuation weight to suppress the extension in noisy speech pauses, which is driven by a robust speech pause detection (SPD), will be introduced in Sec. 4.2.3.

Fig. 2.9a schematically illustrates ST for $\Omega_{\mathrm{M}} = \Omega_{\mathrm{c}}$, given a frequency band specification of $f_{\mathrm{c}} = 3.6$ kHz, i.e., $\Omega_{\mathrm{c}} = 2\pi f_{\mathrm{c}}/f_{\mathrm{s}} = 0.45\pi$. In this case, an IIR anti-aliasing highpass filter with a cut-off frequency equal to f_{c} is required to remove the spectral alias components $|E_{\mathrm{LB}}(e^{j(\Omega+\Omega_{\mathrm{M}})})|/2$ from the LB amplitude spectrum $|E_{\mathrm{LB}}(e^{j\Omega})|$. Furthermore, a modulation gain of $g_{\mathrm{M}} = 2$ allows for an energy adaptation yielding the estimated UB amplitude spectrum $|\hat{E}_{\mathrm{UB}}(e^{j\Omega})| = |E_{\mathrm{LB}}(e^{j(\Omega-\Omega_{\mathrm{M}})})|$. In the general case of $\mathrm{mod}(\Omega_{\mathrm{M}}, 2\pi F_0/f_{\mathrm{s}}) \neq 0$, with $\mathrm{mod}(\,\cdot\,)$ denoting the modulo operator, the harmonic structure is not reconstructed correctly.

Spectral folding (SF) represents a special case of ST for $\Omega_{\mathrm{M}} = \pi$ (Jax, 2002, Sec. 3.3.1). As depicted in Fig. 2.9b, it yields a mirroring of the LB amplitude spectrum $|E_{\mathrm{LB}}(e^{j\Omega})|$. This provokes a spectral gap in the frequency range $\Omega_{\mathrm{c}} \ldots \Omega_{\mathrm{M}} - \Omega_{\mathrm{c}} = 0.45\pi \ldots 0.55\pi$. A correct reconstruction of the harmonic structure is neither accomplished. However, aliasing effects are completely prevented in the LB frequency range. The estimated UB amplitude spectrum $|\hat{E}_{\mathrm{UB}}(e^{j\Omega})|$ arises from a constructive superposition of $|E_{\mathrm{LB}}(e^{j(\Omega-\Omega_{\mathrm{M}})})|/2$ and $|E_{\mathrm{LB}}(e^{j(\Omega+\Omega_{\mathrm{M}})})|/2$. Hence, there is no need for an anti-aliasing filter and the modulation gain

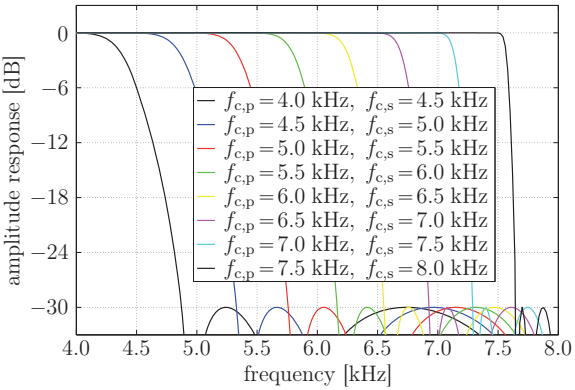


Figure 2.10: Amplitude response of the IIR lowpass postfilter with a variable transition range specified by the passband and stopband cut-off frequencies $f_{c,p}$ and $f_{c,s}$, respectively.

results in $g_M = 1$. Thus, (2.82) is simplified to (Jax, 2002, Eq. (3.4))

$$\hat{e}_{UB}(n) = e_{LB}(n) \cdot \cos(\pi n) \cdot g_{UB} = e_{LB}(n) \cdot (-1)^n \cdot g_{UB}. \tag{2.84}$$

Potential artifacts, which originate from DC components that are spectrally folded to the Nyquist frequency of 8 kHz, can be efficiently suppressed by the subsequent ABE postfilter.

2.2.10 ABE Postfiltering

Compared to (Jax, 2002, Sec. 2.3), the upper cut-off frequency of the estimated WB output speech $\hat{s}_{WB}(n)$ can be optionally adapted by means of a variable lowpass postfilter in the final block of Fig. 2.7. Thus, the degree of bandwidth extension can be controlled.

Computational complexity and algorithmic delay are saved by using a fifth-order elliptic IIR filter design with a stopband attenuation of 30 dB and a passband ripple of 0.01 dB (Oppenheim and Schafer, 1989, App. B.3). To allow for a reasonable variation of the upper cut-off frequency between $4 \ldots 8$ kHz, eight filters have been designed with a 0.5 kHz-wide transition range being shifted in steps of 0.5 kHz. They are specified by the passband and stopband cut-off frequencies $f_{c,p}$ and $f_{c,s}$, respectively. Fig. 2.10 illustrates the amplitude response of the IIR lowpass postfilter focusing on the variable transition range.

The transition range specification $f_{c,p}$ and $f_{c,s}$ of the postfilter can be adapted to the level of background noise. This also applies to the static attenuation weight g_{UB} of the UB residual signal in Sec. 2.2.9. Thus, a more aggressive extension may be allowed in case of noisy environments. Based on the assumption that the noise level is somehow correlated with the speed of a cruising vehicle, the aggressiveness of the ABE should be automatically controlled in automotive applications via the driving speed information on the CAN bus.

2.3 Summary

In this chapter, a complete ABE framework exploiting phonetic *a priori* knowledge is formulated as a further development from the state-of-the-art approach of Jax (2002). Both ABE training and processing are thereby phonetically supported in terms of frame-wise phoneme class labels. This innovation aims at improving speech intelligibility and quality by a reduction of artifacts. They typically arise from ABE due to the confusion of critical phonemes, as mentioned in the preceding chapter. By means of the phonetic support, a *supervised* CB training is created, from which the subsequently trained LDA and HMM can implicitly benefit. Within ABE processing, the phonetic information is integrated into the HMM decoder. It thereby serves as an additional observation along with the extracted features. This yields a modification of the observation likelihoods in terms of a novel phoneme class probability matrix. Besides the exploitation of phonetic *a priori* knowledge, further algorithmic innovations relating to (Jax, 2002) are introduced, such as a frame conversion using non-rectangular windowing with window overlap, a VA-based optimal state sequence decoder, a cepstral smoothing strategy, as well as an additional control over the UB energy and cut-off frequency.

Based on these algorithmic fundamentals, the two following chapters categorize the wide field of application into the main practice-relevant use cases. On the one hand, Chap. 3 deals with a human-to-machine ABE application, which employs an *offline* ABE for the training of WB telephony ASR systems. On the other hand, a human-to-human ABE application is treated in Chap. 4 making use of an *online* ABE for the enhancement of NB telephone speech services. In fact, phonetic *a priori* knowledge in support of the ABE processing is or can be made available only for applications without online requirements. Obviously, this condition is fulfilled in Chap. 3 but not in Chap. 4.

Chapter 3

Human-to-Machine ABE Application: Offline ABE for Training of WB Telephony ASR Systems

In the previous chapter, a complete ABE framework exploiting phonetic *a priori* knowledge in support of both training and processing has been presented. This chapter makes use of it in a human-to-machine ABE application for telephony-based IVR systems. They intend to automatically recognize human speech over the telephone by employing an ASR.

For one century, ASR has been an increasingly important field of research in human-to-machine interaction. A toy called 'Radio Rex' was commercially developed in 1911 as the first ASR machine (Cohen et al., 2004). In the fifties, the tasks were still limited to a recognition of digits or vowels (Marill, 1961). As the computing power has been enormously increased over the past decades, more and more demanding recognition tasks, such as a naturally spoken conversation or dictation, have been facilitated by the use of sophisticated HMMs and ANNs (Huang et al., 2001). According to Church and Mercer (1993), the availability of large speech databases for training purposes is particularly important. Due to the strong impact of the acoustic bandwidth on recognition performance as reported in Sec. 1.1, ASR tasks with large vocabulary and noisy environments generally operate at 16 kHz sampling rate.

However, the support of HD telephony services poses a problem for IVR systems. Despite the large number of available speech corpora originating from conventional telephony, the required ASR training suffers from the lack of WB telephone speech databases. This problem will be tackled by upgrading existing NB telephone speech data via ABE. Since an ASR training does not require online capabilities, the phoneme class labels needed for ABE can be made available offline via phonetic transcription, i.e., either manually (by humans) or

automatically (by forced Viterbi alignment), unless they are already provided along with the respective database. Instead of improving speech intelligibility and quality from a human point of view, such a database extension aims at increasing ASR performance. Compared to the recording of new WB telephone speech corpora, it is less expensive in terms of time and costs.

In this chapter, preliminary NB and WB phoneme recognition experiments identify those phonemes, which benefit most from an acoustic bandwidth increase, to optimize the offline ABE phoneme-specifically. Based on these findings, phonetically motivated CBs are designed for ABE. To counteract potential phonetic over-representations caused by such CB designs, the pre-trained state transition probabilities are modified. Further introduced modifications aim at preventing sparse data problems and suppressing temporally smeared offsets. Finally, large-vocabulary ASR experiments are conducted with a limited amount of simulated WB telephone speech training data to investigate the ABE abilities for a realistic scenario in practice.

3.1 Preliminary Phoneme Recognition Experiments

Before tackling large-vocabulary ASR tasks, this section investigates the phoneme-specific ASR performance by means of preliminary phoneme recognition experiments. From this investigation, the next section derives a phonetically motivated CB design for ABE. In this context, the performance dependency on acoustic bandwidth is of particular interest. Furthermore, the influence of telephone speech transmission characteristics is important for the given telephony-based IVR application. For this purpose, two new derivatives of the well-known TIMIT corpus (Garofolo et al., 1993) have been created comprising WB and NB telephone speech data, respectively. On the one hand, *WTIMIT* has been derived from the TIMIT corpus by a transmission over T-Mobile's 3G mobile network in The Hague, The Netherlands, employing the WB AMR speech codec at the commonly used bit rate of 12.65 kbps (Bauer et al., 2010d,c). It has been released for distribution by the Linguistic Data Consortium (LDC) in 2010 (Bauer and Fingscheidt, 2010). On the other hand, the complementary *NB-TIMIT* has been correspondingly derived from the decimated TIMIT corpus by a transmission over T-Mobile's 3G mobile network in Braunschweig, Germany, employing the NB AMR speech codec at the commonly used bit rate of 12.2 kbps (Bauer et al., 2010d,c).

Covering about 5.5 h of WB telephone speech data, the WTIMIT corpus is too small for the purpose of large-vocabulary ASR training. Due to its phonetically balanced speech material and manually transcribed phoneme labels, however, it is well suited for more restricted phoneme recognition evaluations (Lee and Hon, 1989). Hence, a WB telephony phoneme

recognition experiment is carried out on WTIMIT. In order to identify phonemes that suffer from a limited acoustic bandwidth, another phoneme recognition experiment is performed on the NB-TIMIT corpus providing the NB telephone speech data corresponding to WTIMIT.

In the following, the experimental setup is briefly described, focusing on the most important aspects of the employed phoneme recognizer that is contained in the Hidden Markov Model Toolkit (HTK) (Young et al., 2006). Subsequently, the experimental results are evaluated by emphasizing the dependency of the recognition performance on acoustic bandwidth. Please note that the remainder of this section is largely based on (Bauer et al., 2010d,c).

3.1.1 Experimental Setup

In the acoustic front end of the phoneme recognizer the feature extraction takes place. At first, a DC offset compensation is applied to the speech files (Young et al., 2006, Sec. 5.2). Subsequently, 12 mel-frequency cepstral coefficients (MFCCs) and a log-energy parameter are extracted frame by frame using a Hamming window of 25 ms length and a frame shift of 10 ms (Young et al., 2006, Sec. 3.1.5). In the WB and NB case, 26 and 23 triangular bandpass filters being equally spaced along the mel scale are used for the involved filterbank analysis (Young et al., 2006, Sec. 5.4), respectively. The resulting static features are subject to a file-based cepstral mean normalization (CMN) and log-energy normalization (Young et al., 2006, Sec. 5.6 and Sec. 5.8). Finally, 13 Δ- and 13 $\Delta\Delta$-features are derived from them (Young et al., 2006, Sec. 5.9). Hence, the composite feature vectors reveal a dimension of 39 each.

The acoustic model training of the phoneme recognizer is performed on the predefined training set of the respective TIMIT derivative. Based on the selected TIMIT phone alphabet in (Lee and Hon, 1989), 48 monophone HMMs are trained (Young et al., 2006, Sec. 3.2). Each HMM contains three states with a left-to-right topology and 16 Gaussian mixtures per state. For initialization purposes, a flat start is used with an initial HMM prototype (Young et al., 2006, Sec. 8.3). The *global speech* mean vector and diagonal covariance matrix are thereby assigned to all Gaussian distributions.

The phoneme recognition engine operates on the predefined test set of the respective TIMIT derivative. Instead of a language model, it simply employs a context-independent grammar in terms of a monophone network (Lee and Hon, 1989, Fig. 3(a)). When evaluating the phoneme recognition performance, successive instances of silence are merged. Furthermore, similar phones are clustered according to (Lee and Hon, 1989), yielding a reduced TIMIT phone set of 39 phoneme-like classes.

Speech dataset	Front end	*PER* [%]	Relative increase of *PER* [%]
WTIMIT	WB	39.57	$\frac{40.78-39.57}{39.57} \cdot 100 \approx 3.06$
NB-TIMIT	NB	40.78	

Table 3.1: PER results of HTK-based phoneme recognition on the WTIMIT and NB-TIMIT corpora for a reduced TIMIT phone set after (Lee and Hon, 1989).

3.1.2 Experimental Results

The phoneme recognition performance is evaluated in terms of the commonly used phoneme error rate (PER) measure (Lopes and Perdigão, 2011)

$$PER = \frac{S_{\text{phn}} + I_{\text{phn}} + D_{\text{phn}}}{N_{\text{phn}}} \cdot 100\ \%, \tag{3.1}$$

with S_{phn}, I_{phn}, and D_{phn} denoting the numbers of *wrongly* substituted, inserted, and deleted phonemes, respectively, in addition to the *true* phoneme number N_{phn}. Tab. 3.1 shows the PER results of both phoneme recognition experiments. The NB phoneme recognizer reveals a PER of 40.78 %, which is expectedly higher than the PER of 39.57 % attained by the WB phoneme recognizer (Bauer et al., 2010d,c). This corresponds to a relative PER increase of about 3.06 %. The reason for such a moderate bandwidth dependency may be related to the rather restricted phoneme recognition task compared with more demanding large-vocabulary ASR tasks.

However, the observed *overall* behavior does not consistently apply to the single phonemes. Fig. 3.1 depicts the phoneme-specific PER results of the NB relative to the WB phoneme recognizer for the reduced TIMIT phone set after (Lee and Hon, 1989). Interestingly, the recognition of some phonemes even benefits from a NB telephone speech transmission (phonemes to the right-hand side of Fig. 3.1), whereas for others it hardly depends on the bandwidth (e.g., /m/, /sh/, /ey/, /g/, and /jh/). The mean of the phoneme-specific relative PERs amounts to 3.91 %. In contrast to the relative PER increase of 3.06 % given in Tab. 3.1, it does not take into account the different phoneme-specific occurrences and therefore treats each phoneme equally. The importance of the fricatives /s/ and /z/, which has already been stated in the context of ABE in Sec. 1.3, is explicitly demonstrated in Fig. 3.1. Due to the bandwidth limitation, a huge PER increase of 30 % or more relative to a recognition on WB telephone speech is provoked. Hence, the fricatives /s/ and /z/ turn out to play the most important role in designing the phonetically motivated CB for ABE.

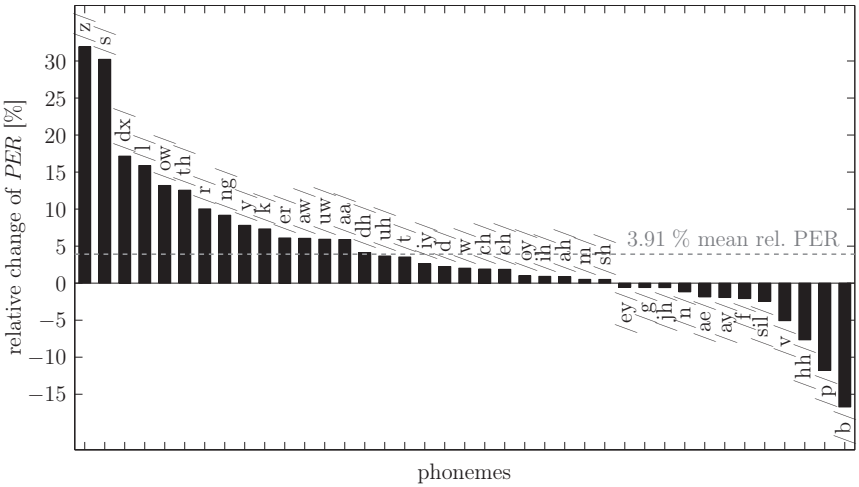


Figure 3.1: Phoneme-specific PER results of the NB relative to the WB phoneme recognizer for the reduced TIMIT phone set after (Lee and Hon, 1989).

3.2 Phonetically Motivated CB Design for ABE

Based on the general concept of *supervised* CB training described in Sec. 2.1.2, the specific phoneme class mapping to single or multiple phonemes is defined in this section via a phonetically motivated CB design. To vary the influence of the phonetic *a priori* knowledge, several ABE CBs are designed with a differing number of phoneme classes. In general parlance, the phonetic influence is increased, when spending more phoneme classes, and vice versa. Before being able to finally train the phoneme-class-specific sub-CBs by means of individual LBG algorithms according to Fig. 2.4, the respective sub-CB size still needs to be determined. Subsequently, some of the most relevant CB designs are exemplarily presented referring to (J. Abel, 2013). The number of defined phoneme classes thereby decreases strictly monotonically.

3.2.1 CB with Multiple Phoneme Classes

37 Phoneme Classes

The first CB design aims at providing a high number of phoneme classes. Of course, this quantity is restricted by the size of the employed phoneme alphabet. Tab. 3.2 shows the designed CB using the phoneme alphabet in Tab. A.1. Inspired by Lee and Hon (1989), some similar phonemes are put into the same phoneme class. Hence, the number of phoneme

φ	Phonemes	$N_{\mathcal{S}}^{(\varphi)}$	φ	Phonemes	$N_{\mathcal{S}}^{(\varphi)}$
0	/si/	32	19	/l/	4
1	/s/	16	20	/j/	4
2	/t/	16	21	/6/	4
3	/n/	16	22	/aI/	4
4	/a/, /a:/, /a∼/	16	23	/NS/	4
5	/VN/	16	24	/S/, /Z/	2
6	/f/	8	25	/x/	2
7	/m/	8	26	/h/	2
8	/I/, /i:/	8	27	/p/	2
9	/U/, /u:/	8	28	/b/	2
10	/O/, /o:/, /o∼/	8	29	/g/	2
11	/@/	8	30	/N/	2
12	/?/	8	31	/Y/, /y:/	2
13	/z/	4	32	/9/, /2:/, /9∼/	2
14	/v/, /w/	4	33	/E/, /E:/	2
15	/C/	4	34	/e:/, /e∼/	1
16	/r/	4	35	/aU/	1
17	/d/	4	36	/OI/	1
18	/k/	4			

Table 3.2: CB design with $N_{\mathcal{P}} = 37$ phoneme classes φ, each with $N_{\mathcal{S}}^{(\varphi)}$ CB entries, and a total size of $N_{\mathcal{S}} = 235$.

classes results in $N_{\mathcal{P}} = 37$. The phoneme class labels φ are incremented in descending order of the sub-CB size $N_{\mathcal{S}}^{(\varphi)}$.

The size $N_{\mathcal{S}}^{(\varphi)}$ of the respective sub-CBs $\mathcal{C}^{(\varphi)}$ has been determined dynamically under the constraint that a sufficient amount of training speech frames per HMM state must be available to avoid sparse data problems. As a rule of thumb, each Gaussian mixture parameter should be trained at least by a minimum of 50 instances (Fingscheidt, 2014, Sec. 6.3). Hence, $50 \cdot (2 \cdot M \cdot d) = 50 \cdot (2 \cdot 8 \cdot 10) = 8000$ instances are required for the complete GMM training of a single HMM state j. This empirical formula takes into account a d-dimensional mean vector $\boldsymbol{\mu}_{j,m}$ and a diagonal $d \times d$ covariance matrix $\boldsymbol{\Sigma}_{j,m}$ for each mixture component m, whereas the scalar weight $\rho_{j,m}$ is neglected here for the sake of convenience. Thus, the sub-CB size $N_{\mathcal{S}}^{(\varphi)}$ is derived from the ratio between the number of available training instances for phoneme class φ and the required number of instances per state (i.e., 8000). To allow for the binary splitting steps of the involved LBG algorithm, the resulting value just

φ	Phonemes	$N_{\mathcal{S}}^{(\varphi)}$
0	/s/, /z/	8
1	/S/, /Z/	8
2	/f/, /v/, /w/	8
3	/C/, /x/, /j/	8
4	/h/, /VN/, /NS/	8
5	/si/	4
6	/p/, /b/, /t/, /d/, /k/, /g/, /?/	8
7	/r/, /m/, /n/, /N/, /l/, /I/, /i:/, /Y/, /y:/, /9/, /2:/, /9∼, /E/, /E:/, /e∼/, /e:/, /U/, /u:/, /O/, /o∼/, /o:/, /a/, /a∼/, /a:/, /6/, /@/, /aI/, /aU/, /OI/	8

Table 3.3: CB design with $N_{\mathcal{P}} = 8$ phoneme classes φ, each with $N_{\mathcal{S}}^{(\varphi)}$ CB entries, and a total size of $N_{\mathcal{S}} = 60$.

needs to be rounded down to the next smaller power of two. The total size of the final CB $\mathcal{C} = \{\mathcal{C}^{(\varphi)} \mid \varphi \in \{0, 1, \ldots, 36\}\} = \{\mathbf{c}^{(i)} \mid i \in \mathcal{S}\}$ results in $N_{\mathcal{S}} = \sum_{\varphi=0}^{36} N_{\mathcal{S}}^{(\varphi)} = 235$ and thereby determines implicitly the number of HMM states.

Eight Phoneme Classes

The second CB design is based on the assumption that not each phoneme is equally important for ABE, as already stated in Sec. 1.3. Hence, the number of phoneme classes is reduced by grouping related phonemes according to their relevance. The more relevant particular phonemes are, the less of them are grouped within a phoneme class. Tab. 3.3 shows the designed CB comprising $N_{\mathcal{P}} = 8$ phoneme classes.

Due to the importance of fricatives for ABE, they are involved in the phoneme classes $\varphi \in \{0, 1, \ldots, 4\}$. The most important fricatives are assumed to be the unvoiced/voiced counterparts /s/ and /z/, /S/ and /Z/, as well as /f/ and /v/. They are therefore put into the distinct phoneme classes $\varphi \in \{0, 1, 2\}$. Fricative /w/ is thereby merged with /f/ and /v/, because of its related place of articulation. For the same reason, sonorant consonant /j/ is folded with the fricatives /C/ and /x/ in phoneme class $\varphi = 3$. As the pronunciation of fricative /h/ is similar to breath sounds, it is put together in phoneme class $\varphi = 4$ with the distorted speech pauses labeled by /VN/ and /NS/. In contrast, the silence label /si/ that represents undistorted speech pauses is assigned to the single phoneme class $\varphi = 5$. All unvoiced and voiced plosives /p/, /t/, /k/ and /b/, /d/, /g/, respectively, as well as the glottal stop /?/ are further combined in phoneme class $\varphi = 6$, since they are supposed to be not that relevant for ABE as the aforementioned fricatives. Even less important than the plosives are assumed to be the remaining 29 phonemes, which consist of the last fricative

φ	Phonemes	$N_{\mathcal{S}}^{(\varphi)}$
0	/s/, /z/	8 (out of 64)
1	/S/, /Z/	8
2	/f/, /v/	8
3	/si/	8
4	/C/, /x/, /h/, /w/, /p/, /b/, /t/, /d/, /k/, /g/, /?/, /j/, /r/, /m/, /n/, /N/, /l/, /I/, /i:/, /Y/, /y:/, /9/, /2:/, /9∼, /E/, /E:/, /e∼/, /e:/, /U/, /u:/, /O/, /o∼/, /o:/, /a/, /a∼/, /a:/, /6/, /@/, /aI/, /aU/, /OI/, /VN/, /NS/	32

Table 3.4: CB design with $N_{\mathcal{P}} = 5$ phoneme classes φ, each with $N_{\mathcal{S}}^{(\varphi)}$ CB entries, and a total size of $N_{\mathcal{S}} = 64$.

/r/, the sonorant consonants /m/, /n/, /N/, and /l/, as well as all 21 vowels and three diphthongs. They are finally clustered in phoneme class $\varphi = 7$.

As the number of available training instances for a particular phoneme class does not necessarily correlate with its ABE relevance, the respective sub-CB size $N_{\mathcal{S}}^{(\varphi)}$ is defined empirically in contrast to the dynamic derivation of the first CB design. For example, the size of the sub-CB representing the most frequent phoneme label /si/ is therefore reduced from 32 to 4. The remaining sub-CB sizes are set to 8. Thus, the total size of the final CB $\mathcal{C} = \left\{\mathcal{C}^{(\varphi)} \mid \varphi \in \{0, 1, \ldots, 7\}\right\} = \left\{\mathbf{c}^{(i)} \mid i \in \mathcal{S}\right\}$ results in $N_{\mathcal{S}} = \sum_{\varphi=0}^{7} N_{\mathcal{S}}^{(\varphi)} = 60$.

Five Phoneme Classes

The third CB design exclusively focuses on the most relevant phonemes for ABE identified in Sec. 1.3. Hence, the phoneme classes $\varphi \in \{0, 1, 2\}$ are assigned to the unvoiced/voiced counterparts /s/ and /z/, /S/ and /Z/, as well as /f/ and /v/, as shown in Tab. 3.4. In contrast to the second CB design, no specific phoneme classes are dedicated to the less relevant, remaining fricatives or plosives. Due to the fact that undistorted speech pauses may be easily confused with /s/ and /z/ based on NB speech, the silence label /si/ is individually represented by phoneme class $\varphi = 3$. All remaining 43 phonemes are just folded in phoneme class $\varphi = 4$.

The size of sub-CB $\mathcal{C}^{(4)}$ is empirically set to $N_{\mathcal{S}}^{(4)} = 32$ allowing for a sufficient spectral discrimination among the numerously assigned phonemes[1]. Based on some informal ABE listening tests focusing on speech pauses, the sub-CB size for the silence phoneme class is increased in comparison with the second CB design from 4 to 8. This size corresponds also to

[1]Please note that this number can also be obtained by adding the sub-CB sizes that belong to the respective phoneme classes of the second CB design in Tab. 3.3 (i.e., $N_{\mathcal{S}}^{(3)} + N_{\mathcal{S}}^{(4)} + N_{\mathcal{S}}^{(6)} + N_{\mathcal{S}}^{(7)} = 32$).

the remaining sub-CBs. The final CB $\mathcal{C} = \left\{\mathcal{C}^{(\varphi)} \mid \varphi \in \{0, 1, \ldots, 4\}\right\} = \left\{\mathbf{c}^{(i)} \mid i \in \mathcal{S}\right\}$ therefore reveals a total size of $N_{\mathcal{S}} = \sum_{\varphi=0}^{4} N_{\mathcal{S}}^{(\varphi)} = 64$.

Another significant difference compared to the previous CB designs concerns the sub-CB training of the phoneme class representing /s/ and /z/. It was found out in (Bauer et al., 2008) that an underestimation of /s/ and /z/ provokes the typical lisping artifact introduced in Sec. 1.3. However, not only wrong classifications in terms of false /s/- and /z/-rejections are assumed to be mainly responsible for that, but also spectral reconstructions using sub-optimally trained CB representatives. Due to the averaging property of the LBG algorithm, the resulting UB spectral envelopes reveal two problems: They turn out to be spectrally too flat on the one hand and suffer from insufficient energy on the other hand. A modified LBG training was therefore proposed in (Bauer and Fingscheidt, 2009a, Sec. 4.1.2), which can be employed by just replacing the conventional LBG algorithm for the training of sub-CB $\mathcal{C}^{(0)}$ in Fig. 2.4. It thereby operates on those UB cepstral envelope vectors $\mathbf{c}_\ell$ that correspond to the phoneme class $\varphi_\ell = 0$ representing /s/ and /z/.

First of all, an initial sub-CB of size 64 is trained with a conventional LBG algorithm to obtain a variety of UB spectral envelope representatives for /s/ and /z/. To tackle the first problem of spectral flatness, all 64 representatives are sorted in decreasing order by their LSD to the phoneme-class-specific mean UB spectral envelope. These LSDs can be easily computed via (2.69) using the respective LPCC vectors (Jax, 2002, Eq. (4.15)). Those foremost 8 out of 64, which have a zeroth LPCC larger than the one of the overall mean representative, are then selected. Thus, the second problem of insufficient energy is addressed. The resulting UB spectral envelopes finally build the required sub-CB.

Fig. 3.2 compares the UB spectral envelope characteristics between the sub-CB representatives for phoneme class $\varphi = 0$ of Tab. 3.3 and 3.4. For the purpose of comparison, the overall mean representative is illustrated in blue. Being averaged over all /s/- and /z/-instances, it is spectrally flat and reveals only a moderate energy. The 8 conventionally LBG-trained sub-CB representatives of the second CB design are marked in red. Except for the three uppermost red curves, they appear more or less like downwards shifted copies of the overall mean representative. Hence, sharply pronounced /s/- and /z/-sounds can hardly be reconstructed by them. This observation confirms the former assumption in (Bauer et al., 2008) that an underestimation of /s/ and /z/ (lisping artifact) can be also caused by a sub-optimal spectral reconstruction. As expected, the 64 temporary representatives marked in black provide a variety of UB spectral envelopes. Obviously, 8 of them marked by the dashed green curves serve as the sub-CB representatives of the third CB design. Due to their large LSD to the overall mean representative, they appear much less spectrally flat. Furthermore, they are located at a significantly higher energy level. Based on them, an adequate spectral reconstruction of sharply pronounced /s/- and /z/-sounds turned out to be feasible (Bauer

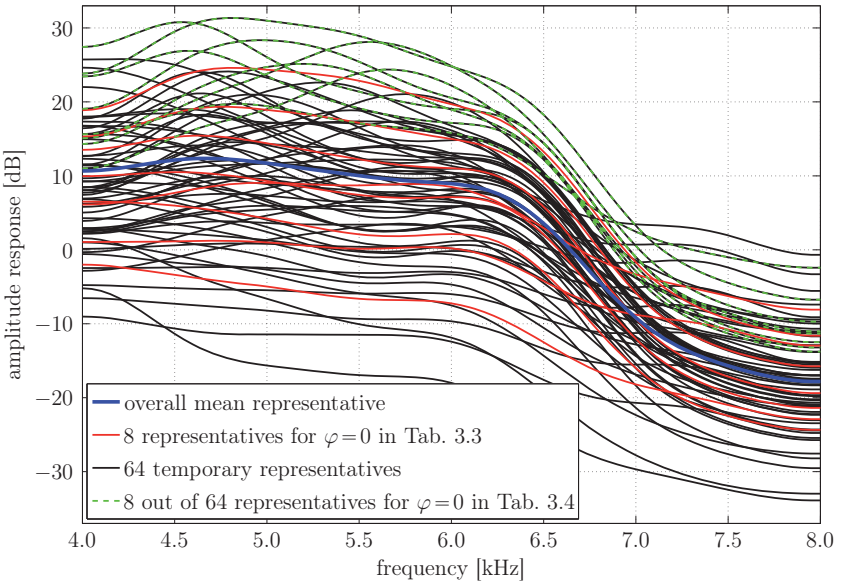


Figure 3.2: UB spectral envelope representatives of variously LBG-trained sub-CBs for the phoneme class with /s/ and /z/ (the respective LPCCs are spectrally converted via (2.71)).

and Fingscheidt, 2009a; Bauer et al., 2010a).

Of course, not all /s/- and /z/-instances are pronounced in a sharp fashion. In fact, the degree of sharpness highly depends on speaker-specific pronunciations, dialects, and languages. In case of a moderate sharpness, a spectral reconstruction relying on the modified LBG training would be unfavorable. Actually, it should be more spectrally flat and less energetic. To solve this problem, a further modification is introduced, but this time with respect to the supervised VQ in Fig. 2.5. At first, those /s/- and /z/-sounds with a zeroth LPCC not exceeding the zero value are detected as moderately sharp. This detection is based on the assumption that sharply pronounced /s/- and /z/-instances have more energy in the UB than in the LB frequency range and therefore reveal – according to (2.17) – a zeroth LPCC larger than zero. In case of a detected moderately sharp /s/- or /z/-pronunciation, a VQ is performed using the modified CB $\mathcal{C} \setminus \mathcal{C}^{(0)}$, i.e., the most appropriate UB spectral envelope is chosen from the representatives of all phoneme-class-specific sub-CBs except for $\mathcal{C}^{(0)}$. In contrast, the VQ of sharply pronounced /s/- and /z/-sounds exclusively involves the sub-CB $\mathcal{C}^{(0)}$, as depicted in Fig. 2.5.

φ	Phonemes	$N_{\mathcal{S}}^{(\varphi)}$
0	/s/, /z/	8 (out of 64)
1	/S/, /Z/, /f/, /v/, /C/, /x/, /h/, /w/, /p/, /b/, /t/, /d/, /k/, /g/, /?/, /j/, /r/, /m/, /n/, /N/, /l/, /I/, /i:/, /Y/, /y:/, /9/, /2:/, /9∼, /E/, /E:/, /e∼/, /e:/, /U/, /u:/, /O/, /o∼/, /o:/, /a/, /a∼/, /a:/, /6/, /@/, /aI/, /aU/, /OI/, /VN/, /NS/, /si/	16

Table 3.5: CB design with $N_{\mathcal{P}} = 2$ phoneme classes φ, each with $N_{\mathcal{S}}^{(\varphi)}$ CB entries, and a total size of $N_{\mathcal{S}} = 24$.

CB with Two Phoneme Classes

The fourth CB design is inspired by the dependency of the phoneme recognition performance on acoustic bandwidth observed in Fig. 3.1. As shown in Tab. 3.5, only two phoneme classes are spent, i.e., $N_{\mathcal{P}} = 2$. While sub-CB $\mathcal{C}^{(0)}$ is dedicated exclusively to /s/ and /z/, sub-CB $\mathcal{C}^{(1)}$ is mapped to all remaining phonemes. This assignment focuses on the outstanding role of the fricatives /s/ and /z/. According to the preceding CB design with five phoneme classes, the modifications of LBG training and supervised VQ are both applied to train the sub-CB $\mathcal{C}^{(0)}$. It therefore reveals a size of $N_{\mathcal{S}}^{(0)} = 8$ (out of 64). The conventionally LBG-trained sub-CB $\mathcal{C}^{(1)}$ turns out to be adequately dimensioned with an empirical size of $N_{\mathcal{S}}^{(1)} = 16$ (Bauer and Fingscheidt, 2009a, Sec. 4.1.2). Thus, the total size of the final CB $\mathcal{C} = \left\{\mathcal{C}^{(0)}, \mathcal{C}^{(1)}\right\} = \left\{\mathbf{c}^{(i)} \mid i \in \mathcal{S}\right\}$ results in $N_{\mathcal{S}} = N_{\mathcal{S}}^{(0)} + N_{\mathcal{S}}^{(1)} = 24$. The resulting UB spectral envelope representatives of both sub-CBs are depicted in Fig. 3.3. As expected, the black curves representing all phonemes except for /s/ and /z/ reveal more spectral flatness as well as less energy than the /s/- and /z/-specific green curves originating from Fig. 3.2.

3.2.2 CB with One Phoneme Class

In contrast to the previous CB designs with multiple phoneme classes, the fifth CB design involves only a single phoneme class representing all of the phonemes, i.e., $N_{\mathcal{P}} = 1$. This reflects the special case of a purely data-driven, unsupervised CB training without phonetic support corresponding to (Jax, 2002, Sec. 6.2). Thus, an exploitation of phonetic *a priori* knowledge within ABE processing is not feasible considering (2.43). The designed CB $\mathcal{C} = \mathcal{C}^{(0)} = \left\{\mathbf{c}^{(i)} \mid i \in \mathcal{S}\right\}$ is trained according to Fig. 2.4 by means of just one conventional LBG algorithm. Taking into account (Jax, 2002, Sec. 6.5.1), the CB size is thereby set to $N_{\mathcal{S}} = 64$.

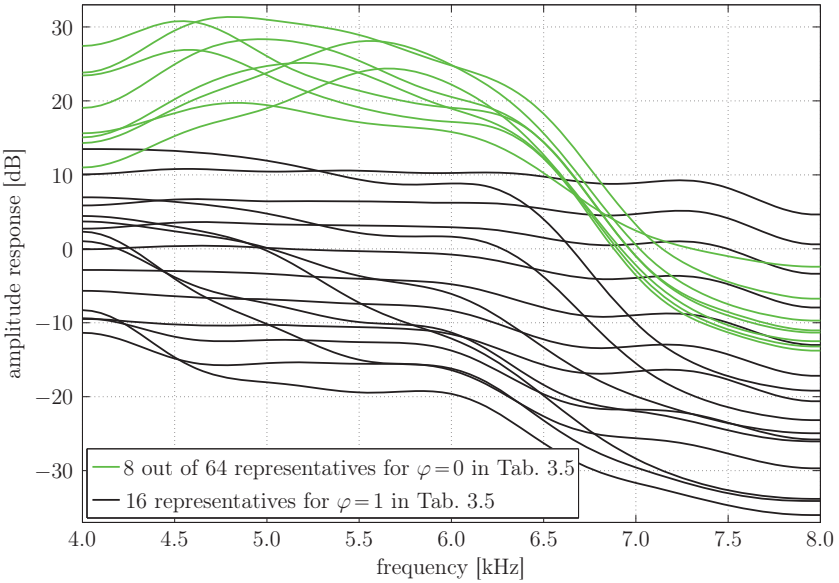


Figure 3.3: UB spectral envelope representatives of the CB design with two phoneme classes.

3.3 Modification of State Transition Probabilities for ABE

The state transition probabilities $a_{i,j} = \mathrm{P}(s_\ell = j | s_{\ell-1} = i)\ \forall\ i, j \in \mathcal{S}$ derived in (2.39) highly depend on the designed CB, which implicitly defines the HMM states. Fig. 3.4 exemplarily illustrates the state transition probability matrix for the phonetically motivated CB design with two phoneme classes. As specified in Tab. 3.5, the first eight states are specifically dedicated to the critical fricatives /s/ and /z/, whereas the subsequent 16 states represent all remaining phonemes. This showcase basically reveals three general types of problems:

1. Single elements of the state transition probability matrix become zero due to insufficient training data (Rabiner and Juang, 1993, Sec. 6.12.4).
2. Because of the dominant main diagonal structure of the state transition probability matrix, the HMM may tend to remain too long in a given state.
3. The over-representation of /s/- and /z/-states provokes temporal smearing effects.

These problems are tackled in Sec. 3.3.1–3.3.3 by introducing particular state transition modifications (Bauer et al., 2014c, Sec. 2.2).

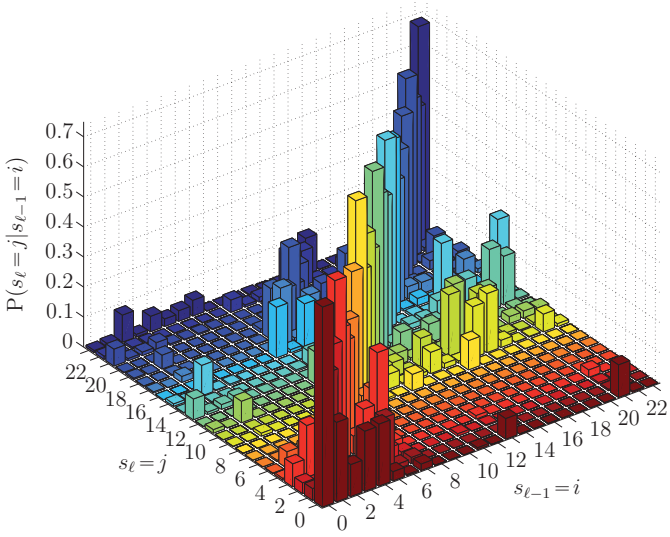


Figure 3.4: Unmodified state transition probability matrix for the CB design in Tab. 3.5.

3.3.1 Smoothing of State Transitions (*)

An HMM is called ergodic, if it permits all mutual combinations of state transitions, i.e., $a_{i,j} > 0 \; \forall \; i, j \in \mathcal{S}$ (Rabiner and Juang, 1993, Sec. 6.3.3). However, zero entries may arise in the joint state histogram $\mathrm{H}(s_\ell = j, s_{\ell-1} = i)$ in case of insufficient training data (Rabiner and Juang, 1993, Sec. 6.12.4). This sparse data problem directly affects the joint state probabilities $\mathrm{P}(s_\ell = j, s_{\ell-1} = i)$ in (2.37) and thereby also the state transition probabilities $\mathrm{P}(s_\ell = j | s_{\ell-1} = i)$ in (2.39).

By means of a state transition smoothing, potential zeros in the joint state histogram are smoothed out by just adding a small value γ (Huang et al., 2001, Sec. 8.4.5)

$$\mathrm{H}^*(s_\ell = j, s_{\ell-1} = i) = \mathrm{H}(s_\ell = j, s_{\ell-1} = i) + \gamma \quad \forall \; i, j \in \mathcal{S}. \tag{3.2}$$

Thus, the modified joint state probabilities are computed according to (2.37)

$$\mathrm{P}^*(s_\ell = j, s_{\ell-1} = i) = \frac{\mathrm{H}^*(s_\ell = j, s_{\ell-1} = i)}{L - 1 + N_{\mathcal{S}}^2 \gamma} \quad \forall \; i, j \in \mathcal{S}, \tag{3.3}$$

with $\sum_{j=0}^{N_{\mathcal{S}}-1} \sum_{i=0}^{N_{\mathcal{S}}-1} \mathrm{H}^*(s_\ell = j, s_{\ell-1} = i) = L - 1 + N_{\mathcal{S}}^2 \gamma$ denoting the total number of state transitions after smoothing. Throughout this work γ is set to one.

3.3.2 Attenuation of State Transitions (**)

To counteract the observation that the HMM, once being in a given state, tends to stay there too long, the main diagonal elements of the joint state probability matrix are attenuated by a constant weighting factor g. Thus, the modified joint state probabilities result in

$$\mathrm{P}^{**}(s_\ell = j, s_{\ell-1} = i) = \frac{\mathrm{P}^{*}(s_\ell = j, s_{\ell-1} = i)}{1 - (1-g) \sum_{\eta=0}^{N_\mathcal{S}-1} \mathrm{P}^{*}(s_\ell = \eta, s_{\ell-1} = \eta)} \cdot \begin{cases} g, & \text{if } i = j \\ 1, & \text{else} \end{cases} \quad \forall\, i, j \in \mathcal{S}. \tag{3.4}$$

For normalization purposes, the denominator takes into account the attenuated elements of the main diagonal. A value of $g = \frac{2}{3}$ turned out to be adequate.

3.3.3 Boosting of State Transitions (***)

Due to the phoneme class assignment of the CB design in Tab. 3.5, /s/- and /z/-sounds are over-represented compared to the remaining phonemes[2]. This over-representation provokes temporally smeared /s/- and /z/-offsets. They can be reduced by emphasizing the transitions from /s/- and /z/-states (i.e., $s_{\ell-1} = i < 8$) back to the others (i.e., $s_\ell = j \geqslant 8$). Following (Sanna and Murroni, 2009, Eq. (3)), the corresponding rectangle of the joint state probability matrix is therefore boosted by an additive gain $\xi > 0$:

$$\mathrm{P}^{***}(s_\ell = j, s_{\ell-1} = i) = \frac{1}{1+128\xi} \cdot \begin{cases} \mathrm{P}^{**}(s_\ell = j, s_{\ell-1} = i) + \xi, & \text{if } i < 8, j \geqslant 8 \\ \mathrm{P}^{**}(s_\ell = j, s_{\ell-1} = i), & \text{else} \end{cases} \quad \forall\, i, j \in \mathcal{S}. \tag{3.5}$$

A value of $\xi = 0.1 \cdot N_\mathcal{S}^{-2}$ (i.e., one-tenth of the uniform distribution) turned out to be suitable. The denominator in (3.5) serves as stochastic constraint involving all $8 \cdot 16 = 128$ elements of the boosted rectangle. For the CB design with five phoneme classes in Tab. 3.4 the boosted rectangle comprises $8 \cdot 56 = 448$ elements. Please note that the state transition boosting is not used for those CB designs proposed in Sec. 3.2, which do not involve the specifically trained phoneme class $\varphi = 0$ with 8 out of 64 /s/- and /z/-representatives.

3.3.4 LSD Performance Evaluation

After the modifications have been applied, the initial state and state transition probabilities are finally recalculated based on the modified joint state probabilities according to (2.38)–(2.39). Fig. 3.5 illustrates the resulting state transition probability matrix, which clearly differs from its unmodified version depicted in Fig. 3.4. While the state transition smoothing

[2]Please note that the fricatives /s/ and /z/ represent about 8 % of all phoneme labels in the phonetically balanced TIMIT corpus.

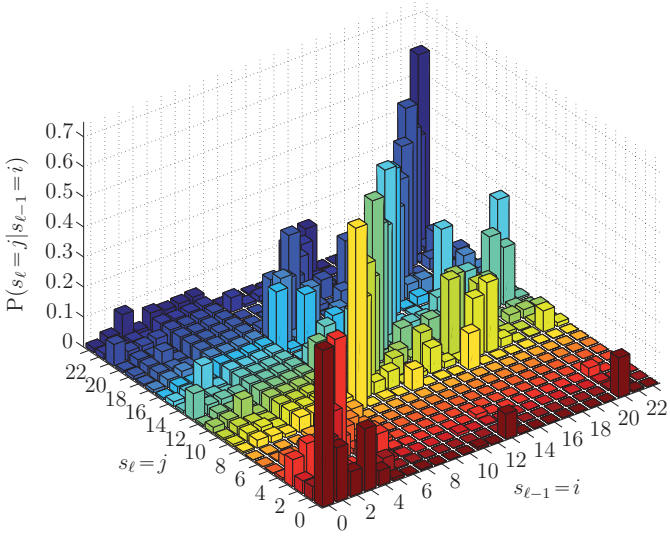


Figure 3.5: Modified state transition probability matrix for the CB design in Tab. 3.5.

is hardly visible due to the small additive value, the attenuation of the main diagonal elements as well as the boosting of the /s/- and /z/-offset rectangle are obvious.

To investigate, whether ABE benefits from the state transition modifications, a reference-based, instrumental performance evaluation is conducted based on the sub-band LSD measure defined in (Pulakka et al., 2008; Katsir et al., 2011). Compared to the LPCC-based LSD measure (Jax, 2002, Eq. (4.15)) exemplarily applied by the cepstral interframe smoothing in (2.69), it derives the frame-wise distortion between a reference and modified speech signal from the respective short-term power spectra $\Phi_{\mathrm{ref},\ell}(k)$ and $\Phi_{\mathrm{mod},\ell}(k)$:

$$LSD_{\mathrm{spec},\ell} = \sqrt{\frac{1}{k_{\mathrm{high}} - k_{\mathrm{low}} + 1} \sum_{k=k_{\mathrm{low}}}^{k_{\mathrm{high}}} \left(10 \log_{10} \frac{\Phi_{\mathrm{ref},\ell}(k)}{\Phi_{\mathrm{mod},\ell}(k)} \right)^2}, \quad 0 \leqslant \ell \leqslant L-1. \tag{3.6}$$

Thus, the LSD can be flexibly specified for a particular spectral band by just restricting the frequency bins to a range $k \in \{k_{\mathrm{low}}, k_{\mathrm{low}}+1, \ldots, k_{\mathrm{high}}\}$. For instance, the LSD specifications in (Pulakka et al., 2008; Katsir et al., 2011) focus on a restricted frequency band ranging from 4 kHz to 7 kHz. Nevertheless, both LSD measures $LSD_{\mathrm{ceps},\ell}$ and $LSD_{\mathrm{spec},\ell}$ are commonly used for speech quality assessment of ABE systems (Pulakka, 2013, Sec. 6.2.1).

The present performance evaluation relies on two sub-band LSD measurements $LSD^{(1)}_{\mathrm{spec},\ell}$ and $LSD^{(2)}_{\mathrm{spec},\ell}$ $\forall\ \ell = 0, 1, \ldots, L-1$. For both of them, all WB speech files originating from the predefined test set of the WTIMIT corpus serve as reference. However, the modified speech signals represent different ABE versions. While the first one makes use of all three state

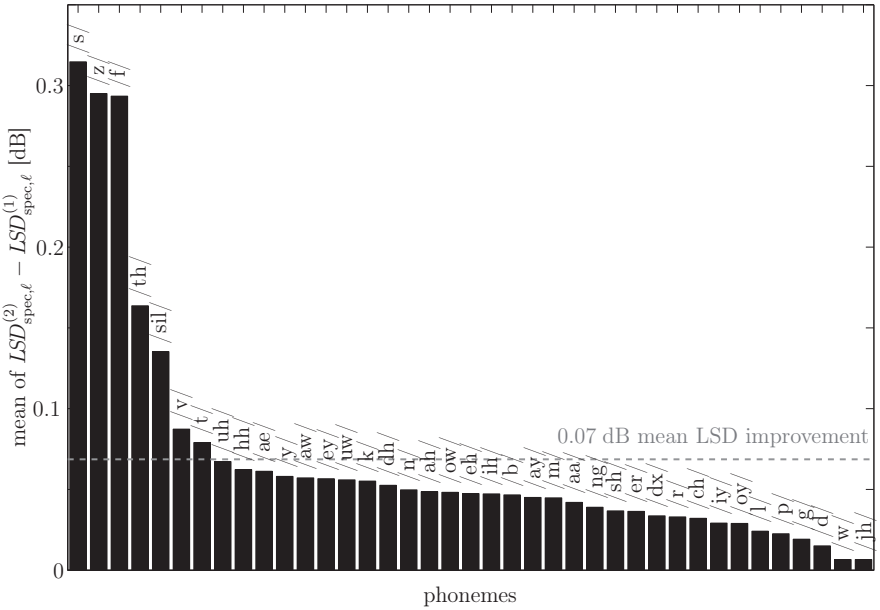


Figure 3.6: Phoneme-specific sub-band LSD improvement attained by an ABE with state transition modifications compared to an unmodified ABE for the reduced TIMIT phone set after (Lee and Hon, 1989).

transition modifications ($LSD^{(1)}_{\mathrm{spec},\ell}$), the second one utilizes none of them ($LSD^{(2)}_{\mathrm{spec},\ell}$). Apart from that, the remaining ABE parametrizations are equal[3]. Moreover, both ABE versions share the same training carried out on the predefined training sets of the NB-TIMIT and WTIMIT corpus. Corresponding to the involved WB reference speech files, the predefined test set of the NB-TIMIT corpus is subject to ABE processing. For both $LSD^{(1)}_{\mathrm{spec},\ell}$ and $LSD^{(2)}_{\mathrm{spec},\ell}$, the sub-band frequency bin edges k_{low} and k_{high} are adapted to a spectral range of $4.4 \ldots 6.4$ kHz. This excludes the spectral gap arising from ABE between about 3.6 kHz and 4.4 kHz because of the SF-based residual signal extension shown in Fig. 2.9b. Furthermore, the WB AMR speech codec at the bit rate of 12.65 kbps used for the creation of the WTIMIT corpus has an internal sampling rate of 12.8 kHz and therefore only provides relevant spectral content up to 6.4 kHz (3GPP TS 26.190, 2001; Bauer et al., 2010d,c).

Fig. 3.6 depicts the phoneme-specific sub-band LSD improvement due to the state transition modifications for ABE. It is characterized by the difference between $LSD^{(2)}_{\mathrm{spec},\ell}$ and

[3]Amongst others, they involve a CB design with two phoneme classes (Tab. 3.5), a phonetic weighting of the state observation likelihoods with $r^{(j)} = 3 \; \forall \; j \in \mathcal{S}$ in (2.43), an FA-based MMSE estimation of the UB cepstral envelope in (2.67), and a residual signal extension by means of SF (Fig. 2.9b).

$LSD^{(1)}_{\text{spec},\ell}$ taking the mean over single phonemes. Obviously, there is no phoneme that suffers from the state transition modifications. In total, a mean LSD improvement of 0.07 dB is achieved. Particularly, the critical fricatives /s/ and /z/ take profits, but also others, like /f/, /th/ as well as the silence label, which tend to be confused with /s/ and /z/. Informal listening tests have also confirmed an improved ABE performance.

3.4 Large-Vocabulary ASR Experiments with ABE

After having initially demonstrated the dependency of the phoneme recognition performance on acoustic bandwidth in Sec. 3.1, this section finally shows the practice-relevant large-vocabulary ASR experiments. The experimental investigations for this purpose have been made in the context of a 2.5 years project funded by the German Research Foundation (DFG) under grant no. FI 1494/4-1 in cooperation with the `European Media Laboratory GmbH` in Heidelberg, Germany, aiming at the transfer of knowledge obtained from the preceding two years basic research project with grant no. FI 1494/2-1. All crucial decisions concerning the experimental design are therefore made from a practice-oriented point of view in agreement with the project partner, who kindly provided his complete ASR training and test processing chain (Fischer and Kunzmann, 2013). It largely relies on the RWTH Aachen University Open Source Speech Recognition Toolkit (Rybach et al., 2001).

The following investigations are largely reported in (Bauer et al., 2014b; J. Abel, 2013). First of all, the setup of the large-vocabulary ASR experiments is described. Before dealing with the ABE-based ASR experiments, practice-relevant ASR baseline results are obtained serving as reference. Finally, the experimental results are discussed to develop a recommendation for offline ABE to upgrade telephone speech databases in bandwidth.

3.4.1 Setup of Large-Vocabulary ASR Experiments

In the following, an overview of the experimental setup is given. To allow for a large-vocabulary recognition task, the ASR training and test process require a large amount of speech data. This also applies to ABE. Hence, an extensive speech database needs to be selected and adequately divided into the respective subsets. Due to the lack of WB telephone speech corpora, realistic telephone speech transmission characteristics have to be simulated by means of a standardized data preprocessing. Thereafter, the most important aspects of the employed ASR framework are briefly described.

Verbmobil subsets	Data amount	Relative portion	Purpose
$\mathbf{ABE_{train}}$	7 h	10 %	ABE training
$\mathbf{ASR_{train}}$	14 h	20 %	ASR training
$\mathbf{ABE_{proc}}$	42 h	60 %	ABE processing, ASR training (!)
$\mathbf{ASR_{eval}}$	7 h	10 %	ASR evaluation

Table 3.6: Shares of data and purposes of the employed Verbmobil subsets.

Definition of Speech Database Subsets

The WTIMIT corpus (Bauer and Fingscheidt, 2010; Bauer et al., 2010d,c) is – to the best of the author's knowledge – the only published WB telephone speech database. Its approximately 5.5 h of American English speech material is by far not sufficient for training of a large-vocabulary recognition task. Furthermore, German was chosen as ASR target language in the DFG project. Hence, the WTIMIT corpus is used only for the more restricted preliminary phoneme recognition evaluations in Sec. 3.1. The German part of the Verbmobil spontaneous speech corpus (Wahlster, 2000) is considered to be large enough for large-vocabulary ASR investigations. It covers 39,935 transcribed dialog turns uttered by 802 speakers resulting in about 70 h of 16 kHz-sampled, uncoded speech data. Being recorded over close-talk and room microphones[4], it still needs to be preprocessed to simulate a realistic telephone speech transmission.

The selected speech data is randomly divided into four speaker-disjoint subsets under the constraint that the gender and age of the speakers are balanced over all subsets. Furthermore, the subset proportions are designed in agreement with the project partner to allow for a practice-oriented scenario. Tab. 3.6 specifies the data amount, relative portion and purpose of the defined Verbmobil subsets. The first subset called $\mathbf{ABE_{train}}$ is spent for ABE training. It includes 7 h (i.e., 10 %) of speech data. The typical use case that there is much less WB than NB telephone speech data available for acoustic model training of telephony-based IVR systems is reflected by the subsets $\mathbf{ASR_{train}}$ and $\mathbf{ABE_{proc}}$. They comprise 14 h (i.e., 20 %) and 42 h (i.e., 60 %) of speech data, respectively. The corresponding data ratio of one third aims at investigating realistic effects of imbalanced data sets for WB and NB ASR training in practice. When applying the ABE processing to subset $\mathbf{ABE_{proc}}$, a huge amount of WB telephone speech data can be acquired in addition to subset $\mathbf{ASR_{train}}$. The last subset denoted as $\mathbf{ASR_{eval}}$ is finally dedicated to ASR evaluation and contains the remaining 7 h (i.e., 10 %) of speech data.

[4]Please note that the small portion of available Verbmobil NB telephone speech recordings has been excluded from this study, as the required WB counterparts do not exist.

Preprocessing of Speech Data

To simulate realistic NB and WB mobile telephone speech, the originally 16 kHz-sampled speech data of the defined Verbmobil subsets is adequately preprocessed taking into account the transmission characteristics of the sending terminal including the speech codec. As the ASR is assumed to operate directly on the decoded signal, the acoustic transmission characteristics of the receiving terminal are excluded. In the NB case, an ABE processing is additionally performed. This leads to the following mobile telephony conditions: **NB**, **ABE**, and **WB**. Any algorithmic delay that may be caused by the involved preprocessing steps is exactly compensated for to preserve the original time alignment.

A flat bandpass filtering to the frequency range of about $0.2 \ldots 3.6$ kHz is applied to model the sending frequency characteristics of NB-capable mobile terminals. It is accomplished by combining the mobile station input (MSIN) highpass filter in (ITU-T G.191, 2009) with a lowpass filter derived from the so-called FLAT1 filter in (ITU-T G.191, 2009). The bandpass-filtered speech signals are decimated to a sampling rate of $f_\mathrm{s}' = 8$ kHz. This implies a high-quality lowpass filtering and subsequent downsampling of factor two. After decimation, a NB mobile telephony call is simulated by applying the NB AMR speech codec at the commonly used bit rate of 12.2 kbps (3GPP TS 26.090, 1999). The resulting **NB** condition directly serves as input for ABE processing. Thus, the ABE-enhanced NB telephony condition **ABE** is obtained. Alternatively, an interpolation to $f_\mathrm{s} = 16$ kHz takes place based on the **NB** condition. This implies an upsampling of factor two and subsequent high-quality lowpass filtering. An interpolated NB telephony condition **NB↑2** is thereby provided for the purpose of the ASR baseline experiments in Sec. 3.4.2.

In contrast, the sending frequency characteristics of WB-capable mobile terminals is modeled by a flat bandpass filtering to the frequency range of about $0.05 \ldots 7.0$ kHz. It is conducted by means of the send-side P.341 weighting filter mask (ITU-T P.341, 2011) in (ITU-T G.191, 2009). To simulate a WB mobile telephony call, the WB AMR speech codec is applied at the commonly used bit rate of 12.65 kbps (3GPP TS 26.190, 2001). In addition to the resulting **WB** condition, a correspondingly decimated WB telephony condition **WB↓2** is created for the ASR baseline experiments in Sec. 3.4.2.

Description of ASR Framework

The employed ASR framework (Fischer and Kunzmann, 2013), which has been kindly provided by the project partner, largely relies on the RWTH Aachen University Open Source Speech Recognition Toolkit (Rybach et al., 2001). In the following, a brief description is given focusing on the acoustic front end, acoustic model training, and speech recognition process (Bauer et al., 2014b, Sec. 3). For further details please refer to (Lööf et al., 2007;

Rybach et al., 2009, 2011).

The acoustic front end of the ASR framework extracts MFCC features using a Hamming window of 25 ms length and a frame shift of 10 ms (Welling et al., 1997). A filterbank analysis is thereby employed consisting of triangular bandpass filters equally spaced along the mel scale. For speech signals sampled at 16 kHz, 16 MFCCs are extracted using 20 bandpass filters. In case of 8 kHz-sampled speech signals, the higher NB frequency components are slightly accentuated by a first-order FIR pre-emphasis filter in advance of the filterbank analysis. Furthermore, only 15 bandpass filters are used and the number of MFCCs is reduced to 12. The resulting MFCC features are subject to a file-based CMN. To capture the temporal dynamics, nine successive feature vectors centered around the current speech frame are concatenated and an LDA transformation is applied reducing the dimension of the composite feature vector to 45 (Lööf et al., 2007, Sec. 2.1).

The acoustic model training of the ASR framework largely follows (Rybach et al., 2009, Sec. 3), however, vocal tract length normalization (VTLN) as well as constrained maximum likelihood linear regression (CMLLR) for speaker adaptive training (SAT) (Rybach et al., 2009, Sec. 3.3) have deliberately not been used in our work to focus more on the acoustic signal properties. For initialization purposes, a flat start is applied to obtain context-independent single-state HMMs. Subsequently, a phonetic decision tree for state tying is trained via classification and regression tree (CART) estimation (Rybach et al., 2009, Sec. 3.1). Two iterations of CART and LDA training are used for the creation of 6000 context-dependent triphone HMMs with a strict left-to-right topology. The transition probabilities are defined independently from the HMM state for the loop, forward, skip, and exit transitions. GMMs with a single, globally pooled diagonal covariance matrix are trained to model the emission probabilities of the HMM states. The GMM parameters are iteratively estimated by means of the well-known EM algorithm. A mixture splitting step between the third and seventh training iteration finally increases the acoustic resolution of the GMM by introducing small perturbations to the mean vectors (Rybach et al., 2009, Fig. 1). This results in a total number of approximately 700,000 45-dimensional GMM densities.

A stochastic trigram language model as well as a pronunciation lexicon organized as a prefix tree have been provided by the project partner to be used together with the trained acoustic models for the speech recognition process (Rybach et al., 2009, Sec. 5). Thus, a time-synchronous beam search algorithm is employed by the speech recognizer (Nolden et al., 2010). On the one hand, the language model is derived from the Verbmobil subset $\mathbf{ASR_{train}}$ by a modified Kneser-Ney smoothing and contains about 100,000 n-grams. To parametrize the weighting exponent of the language model, some preliminary experiments have been carried out. On the other hand, the pronunciation lexicon originates from all defined Verbmobil subsets in Tab. 3.6 to not unnecessarily restrict the large-vocabulary task

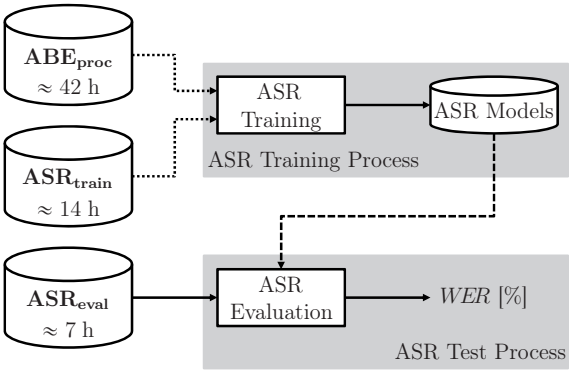


Figure 3.7: Generic block diagram of ASR baseline experiments.

and to prevent out-of-vocabulary effects. It comprises roughly 11,400 words and 12,500 pronunciations.

3.4.2 ASR Baseline Experiments

Based on the defined Verbmobil subsets and the preprocessed telephony conditions, practice-oriented ASR baseline experiments are designed without making use of ABE. They are schematically depicted in terms of the generic block diagram in Fig. 3.7 and serve as reference for the ABE-based ASR experiments in the next section. At the top of Fig. 3.7, the ASR training process takes place to obtain the required HMM-based acoustic models. Either both subsets $\mathbf{ABE_{proc}}$ and $\mathbf{ASR_{train}}$ or only one of them can be used for ASR training. This flexible design is characterized by the dotted signaling paths. At the bottom of Fig. 3.7, the trained ASR models are evaluated via the commonly used WER measure (Huang et al., 2001, Eq. (9.3))

$$WER = \frac{S_{\text{wrd}} + I_{\text{wrd}} + D_{\text{wrd}}}{N_{\text{wrd}}} \cdot 100\ \%, \tag{3.7}$$

with S_{wrd}, I_{wrd}, and D_{wrd} denoting the numbers of *wrongly* substituted, inserted, and deleted words, respectively, in addition to the *true* word number N_{wrd}. For this purpose, the ASR test process employs the subset $\mathbf{ASR_{eval}}$.

Design of ASR Baseline Experiments

The specific design of the ASR baseline experiments is shown in the left part of Tab. 3.7. It assigns the telephony conditions **WB**, **WB↓2**, **NB**, and **NB↑2** to the respective Verbmobil subsets $\mathbf{ABE_{proc}}$, $\mathbf{ASR_{train}}$, and $\mathbf{ASR_{eval}}$. The first two ASR baseline experiments serve as

No.	Verbmobil subsets			*WER* [%]	% *WER* relative to			
	$\mathbf{ABE_{proc}}$	$\mathbf{ASR_{train}}$	$\mathbf{ASR_{eval}}$		*I*	*II*	*III*	*IV* / *V*
I	**WB**	**WB**	**WB**	36.83	±0.0	−6.9	−14.0	−7.8
II	**NB**	**NB**	**NB**	39.58	+7.5	±0.0	−7.6	−0.9
III	–	**WB**	**WB**	42.83	+16.3	+8.2	±0.0	+7.2
IV	**NB**	$\mathbf{WB\downarrow 2}$	$\mathbf{WB\downarrow 2}$	39.95	+8.5	+0.9	−6.7	±0.0
V	$\mathbf{NB\uparrow 2}$	**WB**	**WB**	39.95	+8.5	+0.9	−6.7	±0.0

Table 3.7: Design and results of ASR baseline experiments.

reference for the upper-bound performance in the pure WB and NB case. All three Verbmobil subsets thereby incorporate the telephony condition **WB** and **NB**, respectively. However, experiment *I* and *II* do not meet all of the following constraints, which are considered to be of high practical relevance:

- Due to the lack of WB telephone speech data for ASR training purposes in practice, the larger training subset $\mathbf{ABE_{proc}}$ must do without the **WB** condition. It can be only used for the smaller training subset $\mathbf{ASR_{train}}$ comprising one quarter of training data (this does not apply for experiment *I*).
- Based on the assumption that an *HD Voice* call is being established, WB telephone speech is received by the IVR system and needs to be recognized somehow. This means that the **WB** condition – or at least its decimated version $\mathbf{WB\downarrow 2}$ in case of an additional decimation[5] – must be assigned to the $\mathbf{ASR_{eval}}$ subset (this does not apply for experiment *II*).

In fact, these constraints are met by the remaining ASR baseline experiments. Basically, experiment *III* is based on *I*, except that three quarters of the training data represented by subset $\mathbf{ABE_{proc}}$ are just omitted. However, it still provides a match between the WB telephone speech data involved by the subsets $\mathbf{ASR_{train}}$ and $\mathbf{ASR_{eval}}$. To keep that data match as far as possible and to exploit the large amount of NB telephone speech training data from subset $\mathbf{ABE_{proc}}$, experiment *IV* simply uses the decimated WB telephony condition $\mathbf{WB\downarrow 2}$ for the other subsets $\mathbf{ASR_{train}}$ and $\mathbf{ASR_{eval}}$. The last ASR baseline experiment is based on the third one, i.e., both subsets $\mathbf{ASR_{train}}$ and $\mathbf{ASR_{eval}}$ adopt the **WB** condition, but it additionally includes the interpolated NB telephony condition $\mathbf{NB\uparrow 2}$ in subset $\mathbf{ABE_{proc}}$. Obviously, a mismatch thereby arises from the different speech bandwidths.

[5]Please note that the $\mathbf{WB\downarrow 2}$ condition significantly deviates from the **NB** condition because of the different transmission characteristics and speech codecs simulated by the data preprocessing in Sec. 3.4.1.

Results of ASR Baseline Experiments

The results of the ASR baseline experiments are given in the right part of Tab. 3.7. The resulting WERs range between 36.83 % and 42.83 %. This indicates that the large-vocabulary ASR task is demanding[6]. As expected, the first ASR baseline experiment, which includes the **WB** condition in all subsets, attains the best performance. When omitting three quarter of the involved training data as reflected by experiment *III*, an absolute WER increase of 6.0 % is caused. This explicitly demonstrates the severe problem being created by the limited amount of WB telephone speech data for ASR training. Instead of the absolute WERs, relative WER results are presented in the following to focus in detail on the mutual differences among the ASR baseline experiments.

The WER resulting from experiment *II* is increased by $\frac{39.58-36.83}{36.83} \cdot 100\ \% \approx 7.5\ \%$ relative to the first experiment. This expected degradation can only be explained by the reduced speech bandwidth, as both experiments exploit all 56 h of training data with matching telephony conditions. By reducing the training data to 14 h in the third experiment, the WER is degraded by 16.3 % and 8.2 % relative to experiment *I* and *II*, respectively. The second degradation indicates that the gain in speech bandwidth is overcompensated by the lack of training data. Hence, the last experiments try to complement the training data again at the expense of a bandwidth limitation (experiment *IV*) or data mismatch (experiment *V*).

Due to the increase of training data, the WER resulting from the fourth ASR baseline experiment is reduced by 6.7 % relative to the third experiment, despite the bandwidth limitation. Apart from the small data mismatch between the **NB** and **WB↓2** condition in experiment *IV* producing a slight WER increase of 0.9 % relative to experiment *II*, both NB ASR baseline experiments perform consistently. Surprisingly, the last experiment attains exactly the same performance as the fourth one, although it implies a strong data mismatch due to the different speech bandwidths of the involved **NB↑2** and **WB** condition. In return, an additional decimation prior to the recognition is not required by experiment *V*, which represents an advantage over experiment *IV*.

3.4.3 ABE-Based ASR Experiments

Instead of just interpolating additional NB telephone speech data for WB ASR training, as done in experiment *V*, an ABE can be applied aiming at a reduction of the bandwidth mismatch. The block diagram of the corresponding ABE-based ASR experiments is depicted in Fig. 3.8. It is vertically divided into four processes: ABE training and test as well as ASR

[6]Please note that the WER results reported, e.g., in (Welling et al., 1999; Finke et al., 1997) are considerably lower, however, a direct comparison between the Verbmobil 1996 evaluation set of 43 min length employed there and the test subset $\mathbf{ASR_{eval}}$ with a duration of 7 h used in this work cannot be drawn.

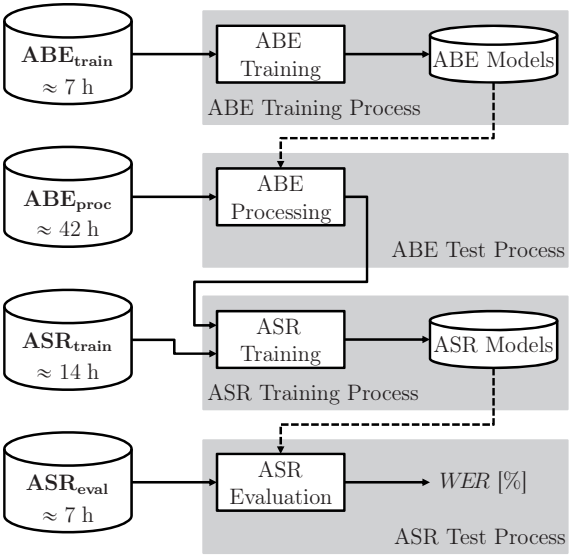


Figure 3.8: Block diagram of ABE-based ASR experiments.

training and test. At the top of Fig. 3.8, the ABE training takes place based on subset $\mathbf{ABE_{train}}$ to obtain the required ABE models[7]. It is followed by an ABE processing of subset $\mathbf{ABE_{proc}}$ using the trained ABE models. Thus, the limited speech bandwidth of the employed **NB** condition is artificially extended. The resulting estimated WB speech data is directly fed into the ASR training process together with subset $\mathbf{ASR_{train}}$ in terms of the **WB** condition. At the bottom of Fig. 3.8, the trained ASR models are finally evaluated by means of subset $\mathbf{ASR_{eval}}$ that also incorporates the **WB** condition.

In the context of this offline ABE application, phonetic *a priori* knowledge in terms of frame-wise phoneme class labels φ_ℓ is used for both ABE training and processing, as explained in Chap. 2. It is provided by the project partner, who carried out a forced Viterbi alignment on the file-based phonetic transcriptions that are already contained in the Verbmobil corpus relying on the extended German speech assessment methods phonetic alphabet (SAMPA) in Tab. A.1. Furthermore, an algorithmic delay of five frames is introduced by deriving the first- and second-order dynamic features from the static ones via (2.31)–(2.32). The interpolation required for ABE processing according to Sec. 2.2.7 makes use of a linear-phase FIR lowpass filter being specified by a stopband cut-off frequency of f_c, a transition

[7]Please note that the NB telephony condition is thereby used along with the unmodified Verbmobil speech data sampled at 16 kHz (direct WB condition). The conceptual decision to perform the ABE training process without using the WB telephony condition accommodates the lack of WB telephone speech data in practice.

WER [%]	% *WER* relative to			
	I	*II*	*III*	*IV* / *V*
38.05	+3.3	−3.9	−11.2	−4.8

Table 3.8: Result of the ASR experiment with upper-bound ABE performance.

range of 100 Hz, a stopband attenuation of 100 dB, and a passband ripple of 0.1 dB. For the residual signal extension described in Sec. 2.2.9, a static weight of $g_{\mathrm{UB,dB}} = -3$ dB is applied to attenuate the estimated UB residual signal. All remaining ABE parameters, which have not been fixed yet, are investigated by the subsequent ASR experiments, except for the optional ABE postfilter described in Sec. 2.2.10 being still neglected here.

ASR Experiment With Upper-Bound ABE Performance

To evaluate the potential of ABE for the given ASR training application, a cheat[8] experiment is conducted reflecting the upper-bound ABE performance. As the estimations of the UB residual signal and WB spectral envelope are considered to be the main issues of ABE processing according to Fig. 2.7, this cheat experiment directly accesses the *original* UB residual signal $e_{\mathrm{UB}}(n)$ and WB LP filter coefficient vector $\mathbf{a}_{\mathrm{WB},\ell}$ instead of using the estimated ones ($\hat{e}_{\mathrm{UB}}(n)$ and $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$). They are obtained by means of an LP analysis based on the original WB speech signal $s_{\mathrm{WB}}(n)$. This actually means that that the UB frequencies are perfectly reconstructed. However, the LB spectrum still reveals a mismatch compared with the original signal due to the applied transmission characteristics and speech codec simulating a NB- instead of a WB-capable terminal.

Tab. 3.8 shows the result of the ASR experiment with upper-bound ABE performance. Obviously, it outperforms all ASR baseline experiments except for the first one having access to the complete training data in terms of the **WB** condition. The resulting performance degradation of 3.3 % relative to experiment *I* can be explained by the aforementioned spectral mismatch in the LB frequency range. Nevertheless, a significant performance gain of 3.9 . . . 11.2 % WER relative to the other experiments *II*-*V* can be observed. Based on the assumption that the ABE performance is lower-bounded by a simple interpolation (**NB↑2**, experiment *V*) and upper-bounded by a perfect UB reconstruction (cheat experiment), a relative WER gap of (only) 4.8 % arises. In the following, this potential is tried to be exploited as far as possible by investigating several impacts on ABE performance.

[8] Throughout this work, the term cheat shall be understood as a trick instead of a deception.

ABE parametrization			*WER* [%]	% *WER* relative to *I*	*II*	*III*	*IV* / *V*
FA	SF	f_c = 4.0 kHz	**39.30**	+6.7	−0.7	−8.2	−1.6
		f_c = 3.6 kHz	39.56	+7.4	−0.1	−7.6	−1.0
	ST	f_c = 3.6 kHz	39.75	+7.9	+0.4	−7.2	−0.5
FBA	SF	f_c = 4.0 kHz	**39.25**	+6.6	−0.8	−8.4	−1.8
		f_c = 3.6 kHz	39.93	+8.4	+0.9	−6.8	−0.1
	ST	f_c = 3.6 kHz	39.76	+8.0	+0.5	−7.2	−0.5
VA	SF	f_c = 4.0 kHz	39.31	+6.7	−0.7	−8.2	−1.6
		f_c = 3.6 kHz	**39.07**	+6.1	−1.3	−8.8	−2.2
	ST	f_c = 3.6 kHz	39.08	+6.1	−1.3	−8.8	−2.2

Table 3.9: Impact of residual signal extension and frequency band specification on decoder-specific ABE performance. The smallest WER for each HMM decoder is marked in bold.

Impact of Residual Signal Extension and Frequency Band Specification on ABE

To analyze the influence of the residual signal extension and frequency band specification on ABE performance, the SF and ST techniques described in Sec. 2.2.9 are compared for a cut-off frequency of $f_c \in \{3.6\,\text{kHz}, 4\,\text{kHz}\}$ specifying the LB and UB frequency range in Sec. 2.1.2. Independent from the choice of f_c, a spectral gap between about 3.6 kHz and 4.4 kHz is caused by SF. In contrast, ST can manage a residual signal extension without interruption for $f_c = 3.6$ kHz by adapting the modulation frequency to $\Omega_M = 2\pi \cdot 3.6\,\text{kHz}/f_s$ and the modulation gain to $g_M = 2$. Of course, this assumes the upper cut-off frequency of the NB telephone speech to be 3.6 kHz, according to Sec. 3.4.1. An 80th-order IIR highpass filter with a cut-off frequency of 3.6 kHz is finally used for ST to prevent aliasing.

The impact of SF and ST in combination with f_c is evaluated for the FA- and FBA-based optimal state decoders using the MMSE estimation (2.67) as well as for the VA-based optimal state sequence decoder using the Viterbi path (2.68). By setting the phonetic weighting ratio defined in (2.46) to $r^{(j)} = 1 \;\forall\; j \in \mathcal{S}$, the phonetic *a priori* knowledge is still neglected. This evaluation furthermore employs the CB design with two phoneme classes in Sec. 3.2.1 as well as all modifications of the state transition probabilities according to Sec. 3.3.

The corresponding evaluation results are given in Tab. 3.9. Obviously, the lower-bound ABE performance represented by experiment *V* is consistently exceeded. All HMM decoders attain the highest performance, when using the SF technique. Hence, it can be concluded that a moderate spectral gap around 4.0 kHz is not critical for ASR training purposes. While the FA- and FBA-based decoders perform better with a frequency band specification of $f_c = 4.0$ kHz, a cut-off frequency of $f_c = 3.6$ kHz turns out to be more suitable for the VA-

ABE parametrization		*WER* [%]	% *WER* relative to *I*	*II*	*III*	*IV* / *V*
FA	$r = 1$	39.30	+6.7	−0.7	−8.2	−1.6
	$r = 3$	**39.21**	+6.5	−0.9	−8.5	−1.9
	$r = 5$	39.42	+7.0	−0.4	−8.0	−1.3
	$r = 7$	39.84	+8.2	+0.7	−7.0	−0.3
FBA	$r = 1$	39.25	+6.6	−0.8	−8.4	−1.8
	$r = 3$	39.24	+6.5	−0.9	−8.4	−1.8
	$r = 5$	**39.19**	+6.4	−1.0	−8.5	−1.9
	$r = 7$	39.29	+6.7	−0.7	−8.3	−1.7
VA	$r = 1$	39.07	+6.1	−1.3	−8.8	−2.2
	$r = 3$	38.95	+5.8	−1.6	−9.1	−2.5
	$r = 5$	**38.83**	+5.4	−1.9	−9.3	−2.8
	$r = 7$	39.28	+6.7	−0.8	−8.3	−1.7

Table 3.10: Impact of phonetic *a priori* knowledge on decoder-specific ABE performance. The smallest WER for each HMM decoder is marked in bold.

based decoder. These ABE parametrizations are kept for the remaining ASR experiments.

Impact of Phonetic *A Priori* Knowledge on ABE

Based on the preceding evaluation, the influence of the phonetic *a priori* knowledge is investigated by varying the phonetic weighting ratio $r^{(j)}$ defined in (2.46). This parameter variation is done independently from the HMM state by fixing $r^{(j)} = r \; \forall \; j \in \mathcal{S}$ with $r \in \{1, 3, 5, 7\}$.

Tab. 3.10 shows the results of this investigation. Obviously, the performance without using phonetic *a priori* knowledge ($r = 1$) can be consistently exceeded by all HMM decoders. While the FA-based HMM decoder performs best at a ratio of $r = 3$, the FBA- and VA-based decoders reveal the highest performance for $r = 5$. When further increasing the phonetic influence and thus trusting more in the phoneme class labels φ_ℓ, the results degrade again. Hence, it can be concluded that phonetic *a priori* knowledge helps to improve the performance only to a certain extent. This observation could be explained by the restricted reliability of the phonetic transcription process, which has been carried out automatically via a forced Viterbi alignment. Apart from that, the VA-based HMM decoder outperforms the FA- and FBA-based HMM decoders for every single ratio. This indicates that the optimal state sequence is superior to the optimal states. For the remaining ASR experiments, the VA-based HMM decoder is therefore employed using a phonetic weighting ratio of $r = 5$.

CB design	*WER* [%]	% *WER* relative to			
		I	*II*	*III*	*IV* / *V*
$N_{\mathcal{P}} = 37$	39.63	+7.6	+0.1	−7.5	−0.8
$N_{\mathcal{P}} = 8$	39.61	+7.5	+0.1	−7.5	−0.9
$N_{\mathcal{P}} = 5$	39.04	+6.0	−1.4	−8.8	−2.3
$N_{\mathcal{P}} = 2$	**38.83**	+5.4	−1.9	−9.3	−2.8
$N_{\mathcal{P}} = 1$	39.02	+5.9	−1.4	−8.9	−2.3

Table 3.11: Impact of CB design on ABE performance using the VA-based HMM decoder. The smallest WER is marked in bold.

Impact of CB Design on ABE

The evaluation of the phonetically motivated CB designs in Sec. 3.2 is based on the previously found ABE parametrizations, i.e., SF with $f_c = 3.6$ kHz as well as VA with $r = 5$. According to (2.43), phonetic *a priori* knowledge is not exploited in case of the CB design with one phoneme class defined in Sec. 3.2.2. Moreover, the boosting modification of the state transition probabilities is only used for the CB designs with two and five phoneme classes, as explained in Sec. 3.3.3.

The corresponding evaluation results are given in Tab. 3.11. Obviously, the CB design with two phoneme classes attains the best performance resulting in a WER of 38.83 %. This yields a WER improvement of 2.8 % relative to the lower-bound ABE performance represented by experiment *V*. Interestingly, the WER degrades consistently, when further increasing the number of phoneme classes. In contrast, the purely data-driven CB design with $N_{\mathcal{P}} = 1$ turns out to perform only slightly worse than the phonetically motivated CB design with $N_{\mathcal{P}} = 2$. In fact, only the phoneme class representing /s/ and /z/ turns out to be beneficial, which confirms the results of the preliminary phoneme recognition experiments in Fig. 3.1.

3.4.4 Discussion

The conducted ASR baseline experiments point out the problem of insufficient WB telephone speech data to train IVR systems supporting *HD Voice* services in practice. This lack of training data provokes in the given showcase a relative WER degradation of 16.3 % (cf. Tab. 3.7, experiment *III* vs. *I*). When additionally making use of NB telephone speech data for WB ASR training, this relative WER degradation is at least reduced to 8.5 % (cf. Tab. 3.7, experiment *IV* / *V* vs. *I*). This indicates that the statement "more data are better data" of Church and Mercer (1993, pp. 18–19) is true. To appropriately adapt the

differing sampling rates, either a decimation prior to the recognition or an interpolation of the added training data is required. Assuming that an interpolation represents the most rudimentary form of ABE, it defines the lower-bound ABE performance. In contrast, a perfect UB reconstruction provides the upper-bound ABE performance. The latter reduces the relative WER degradation to 3.3 % (cf. Tab. 3.8, cheat experiment vs. experiment *I*). The potential for ABE can be expressed by directly relating the upper to the lower ABE performance bound. Thus, a (not very large) relative WER gap of 4.8 % arises (cf. Tab. 3.8, cheat experiment vs. experiment *V*). After the employed ABE has been parametrized with respect to the residual signal extension, frequency band specification, phonetic *a priori* knowledge, HMM decoder, and CB design, more than half of this gap, i.e., 2.8 %, is bridged (cf. Tab. 3.11, best ABE experiment vs. experiment *V*). Moreover, the abovementioned relative WER degradation of 16.3 % due to insufficient training data decreases to only 5.4 % (cf. Tab. 3.11, best ABE experiment vs. experiment *I*). The additional use of NB telephone speech data for WB ASR training via ABE results in a relative WER improvement of 9.3 % (cf. Tab. 3.11, best ABE experiment vs. experiment *III*). Please note that comparable results are achieved by Seltzer and Acero (2007, Fig. 4) assuming a WB-to-NB training data ratio of one quarter. However, their utilized feature- and model-based ABE techniques require modifications of the employed ASR system affecting the feature extraction and acoustic model training, respectively. In contrast, the ABE approach presented in this work just operates at a speech data level prior to the ASR training and is therefore completely independent from the speech recognizer.

3.5 Summary

In this chapter, the introduced ABE framework exploiting phonetic *a priori* knowledge is utilized to upgrade NB telephone speech data for the purpose of WB ASR training. This human-to-machine ABE application tackles the problem of insufficient WB telephony training data in practice. As it does not require any online capabilities, the employed ABE is performed offline. Hence, the phoneme class labels in support of both ABE training and processing are (or just can be made) available. Based on preliminary phoneme recognition experiments, the important role of the fricatives /s/ and /z/ is demonstrated. It is taken into account for a phonetically motivated CB design. By means of specifically trained CB representatives, underestimation artifacts that typically arise from ABE are reduced. However, the involved over-representation of /s/- and /z/-states provokes temporal smearing effects. This problem is, amongst others, addressed by modifications of the state transition probabilities. In the context of a practice-oriented scenario, large-vocabulary ASR experiments finally point out the ability of the further developed ABE for the given application.

After having treated this human-to-machine ABE application, which particularly aims at improving the 'intelligibility' for the purpose of an automatic speech recognizer, the next chapter deals with a human-to-human ABE application. It focuses on the enhancement of telephone speech intelligibility and quality from a human perspective. As the employed ABE thereby needs to be performed online, the phonetic *a priori* knowledge is actually available only for ABE training.

Chapter 4

Human-to-Human ABE Application: Online ABE for Enhancement of NB Telephone Speech Services

This chapter employs the ABE framework developed in Chap. 2 to enhance the speech intelligibility and quality of NB calls from a human point of view. The ABE processing thereby needs to meet the online requirements of the telephone speech services. In contrast to the last chapter, phonetic *a priori* knowledge is therefore only available for the offline ABE training process.

In comparison with HD telephony, NB calls suffer from a reduced speech intelligibility and quality because of their limited acoustic bandwidth, according to Sec. 1.1. Hence, the telephone alphabet is often utilized instinctively to spell words without context information, such as proper names. Particularly, conversations held in a foreign language raise the mental load of the conversational partners. In case of an in-car telephone call, the driver could thereby be distracted considerably. Any driver distraction can immediately impair the driving safety. This problem is specifically addressed in (Bauer et al., 2010b) by integrating ABE into the hands-free system of an automobile. Apart from that, the automotive context is also taken into account for the general investigations on the given human-to-human ABE application presented in this chapter.

At first, the influence of ABE on speech intelligibility is analyzed via subjective listening tests with contextless syllables. This analysis somehow corresponds to the instrumental phoneme recognition experiments in Sec. 3.1. To further optimize the ABE with respect to speech quality, several innovations are introduced focusing basically on the reduction of artifacts. Finally, the ability of ABE for speech quality enhancement is assessed subjectively and instrumentally to find out, how to evaluate ABE systems in practice.

4.1 Preliminary Syllable Articulation Tests with ABE

Subjective listening tests with contextless syllables are conducted to examine the impact of ABE on human speech intelligibility, as inspired by French and Steinberg (1947) (see Sec. 1.1). The syllable articulation tests involve the ABE framework, which has been further developed in Chap. 3 to improve ASR performance. It is trained on the phonetically balanced TIMIT corpus[1] (Garofolo et al., 1993) and employs the phonetically motivated CB design with two phoneme classes in Sec. 3.2.1 as well as all state transition modifications in Sec. 3.3. Due to the missing phonetic *a priori* knowledge for ABE processing, the phonetic weighting ratio (2.46) results in $r^{(j)} = 1 \ \forall \ j \in \mathcal{S}$. To allow for an online-capable ABE processing, algorithmic delay contributions are kept low as far as possible. Hence, the first- and second-order dynamic features are obtained by the 'low-delay' derivation method (2.33)–(2.34). Since FBA and VA typically require additional delay, the FA-based optimal state decoder is applied using the MMSE estimation (2.67). According to Tab. 3.9, an SF-based residual signal extension is therefore selected in combination with a frequency band specification of $f_\mathrm{c} = 4.0$ kHz. By attenuating the estimated UB residual signal following Sec. 2.2.9 with a static weight of $g_{\mathrm{UB,dB}} = -9$ dB and applying the lowpass postfilter with a transition range of $5.5 \ldots 6.0$ kHz defined in Sec. 2.2.10 to the estimated WB output speech, the aggressiveness of the ABE is controlled.

Before describing the setup of the syllable articulation tests, the preprocessing of the employed speech data is explained. An NB, ABE-enhanced, and WB telephone speech transmission including two automotive (near-end) background noise levels is thereby simulated. The main objective of the test results is to find out, whether ABE is capable of improving NB telephone speech intelligibility. Apart from the dependency on speech bandwidth, the influence of language and hearing ability is investigated by involving native and non-native, as well as normal-hearing and hearing-impaired listeners.

4.1.1 Experimental Setup

Syllable articulation is tested via three subjective listening tests using vowel-consonant-vowel (VCV) logatomes. They consist of single onset/offset vowels and single center consonants. The emphasis is thereby put on the center consonants, which represent relevant phonemes in the context of ABE according to Sec. 1.3. Amongst them, the involved subjects have to acoustically recognize the correct ones. To get used to the test setup, the subjects first get an initial instruction and familiarization, where a comfortable sound level can be individually adjusted.

[1] The preprocessing of the NB and WB speech data used for ABE training largely follows Fig. 4.1, however, a noise addition is not included.

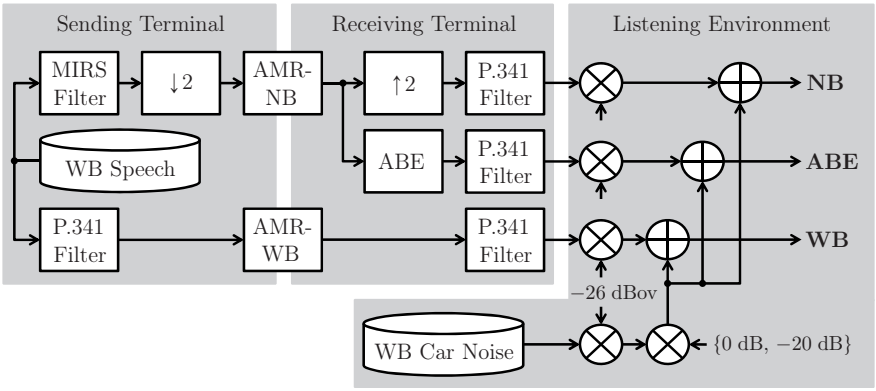


Figure 4.1: Data preprocessing for the syllable articulation tests simulating realistic NB, ABE-enhanced and WB telephony conditions in an automotive near-end listening environment.

Several influences on the syllable articulation are investigated. All listening tests are based on the subsequently explained data preprocessing. It simulates NB, ABE-enhanced and WB telephony conditions in an automotive near-end listening environment at 20 dB and 0 dB SNR. The first listening test serves as reference for the remaining ones, as it deploys normal-hearing listeners, who are native speakers with respect to the pronunciation of the employed syllables. In the second listening test, the language dependency is analyzed by involving normal-hearing listeners, who are *non-native* speakers with respect to the pronunciation of the employed syllables. The impact of hearing ability is finally examined by the third listening test that engages *hearing-impaired* listeners, who are native speakers with respect to the pronunciation of the employed syllables.

Preprocessing of Speech Data

Fig. 4.1 illustrates the preprocessing of the employed WB logatome data. It simulates realistic NB, ABE-enhanced and WB telephony conditions taking into account the transmission characteristics of the sending and receiving terminal including the speech codec. Furthermore, an automotive near-end listening environment is simulated by adding WB car noise with 20 dB and 0 dB SNR. To obtain temporally synchronized telephony conditions, the algorithmic delay of the individual preprocessing steps is exactly compensated for.

In contrast to the preprocessing for the large-vocabulary ASR experiments in Sec. 3.4.1, the send-side modified intermediate reference system (MIRS) weighting filter mask (ITU-T P.830, 1996, Annex D) in (ITU-T G.191, 2009) is utilized instead of the MSIN-/FLAT1-

based flat bandpass filtering to model the sending frequency characteristics of NB-capable mobile terminals. The MIRS-filtered speech signals are subject to a decimation of factor two, followed by an application of the NB AMR speech codec at the commonly used bit rate of 12.2 kbps (3GPP TS 26.090, 1999). On the one hand, the resulting speech signals serve as input for ABE processing. On the other hand, they are interpolated to $f_s = 16$ kHz for the NB telephony condition. The P.341 weighting filter mask (ITU-T P.341, 2011) in (ITU-T G.191, 2009) is employed to model the receiving frequency characteristics of WB-capable mobile terminals. This allows for a fair comparison with the WB telephony condition, while a receive-side MIRS weighting would unnecessarily degrade the syllable articulation of the NB and particularly the ABE-enhanced telephony condition.

In case of the WB telephony condition, the P.341 weighting filter mask (ITU-T P.341, 2011) in (ITU-T G.191, 2009) is utilized twice to model both the sending and receiving frequency characteristics of WB-capable mobile terminals. In between, the WB AMR speech codec is applied at the commonly used bit rate of 12.65 kbps (3GPP TS 26.190, 2001).

To simulate an automotive near-end listening environment, the resulting P.341-filtered speech signals of all telephony conditions are superposed by WB car noise according to (ITU-T P.56, 2011). For this purpose, the active speech levels are consistently adjusted via (ITU-T G.191, 2009) to -26 dBov, while the car noise is scaled via (ITU-T G.191, 2009) to root mean square (RMS) levels of -26 dBov and -46 dBov. A superposition of the corresponding speech and noise signals finally provides the telephony conditions **NB**, **ABE**, and **WB** with SNRs of 0 dB and 20 dB.

First Test with Native, Normal-Hearing Listeners at 20 dB and 0 dB SNR

The first subjective syllable articulation test is performed in our own laboratories by means of an `RME Fireface 400` external sound card, high-quality `AKG K 271 MK II` headphones, and a laptop using a `MATLAB GUI` software (Bauer et al., 2013). It is based on contextless VCV logatomes following the German SAMPA. They are pronounced in German by two male and two female native speakers and originate from (Bellanova et al., 2010, 2011, 2012).

The VCV combinations included in the test combine the single onset/offset vowels /a/, /I/ and /U/ with the single center consonants /s/, /f/, /S/, /x/ and /C/. The latter ones characterize those fricatives that turn out to be the most relevant unvoiced German phonemes for ABE following Sec. 1.3. Please note that /x/ and /C/ are allophones of a single phoneme (the purely German 'ch' sound) and thus denote actually the same center consonant. The velar fricative /x/ (as in 'Nacht' or 'Flucht') is thereby paired only with the open front vowel /a/ and close back vowel /U/, while the palatal fricative /C/ (as in 'Licht') is paired only with the close front vowel /I/ (Wiese, 2010). The remaining fricatives /s/,

/f/, and /S/ are paired with all vowels /a/, /I/ and /U/. This results in a total of 12 VCV combinations.

Taking into account the 12 VCV combinations from all four speakers as well as the three preprocessed[2] telephony conditions **NB**, **ABE**, and **WB**, 144 logatome files are available for the test. Every listener has to recognize the correct center consonants of all these logatome test files in an individually randomized order. For both SNRs, a separate test is conducted, each with 12 predominantly non-expert German listeners having a mean age below 30 years and a normal hearing ability. Hence, all of them can be considered as native (with respect to the German syllable pronunciations) and normal-hearing. The audio samples are presented to the listeners in a diotic manner over the headphones.

Second Test with *Non-native*, Normal-Hearing Listeners at 20 dB and 0 dB SNR

The setup of the second subjective syllable articulation test is exactly based on the first one, except for the following deviations (Bauer et al., 2010b).

In contrast to the first test, contextless VCV logatomes pronounced in British English by eight male and eight female native speakers are employed following the English SAMPA. They originate from the training subset of the Interspeech 2008 Consonant Challenge corpus (Cooke and Scharenborg, 2008).

This time, 24 VCV combinations are included in the test combining the single onset/offset vowels /V/ (corresponding to the German /a/), /I/, and /U/ with the unvoiced fricatives /s/, /f/, /S/, and /T/ as well as their voiced counterparts /z/, /v/, /Z/, and /D/. Besides the addition of the corresponding voiced fricatives, the exclusively German allophones /x/ and /C/ are replaced by /T/ (and /D/) denoting the purely English unvoiced (and voiced) 'th' sound. The fricatives characterizing the single center consonants turn out to be the most relevant English phonemes for ABE following Sec. 1.3. This results in a total of 24 VCV combinations.

Taking into account the 24 VCV combinations from all 16 speakers as well as the three preprocessed[3] telephony conditions **NB**, **ABE**, and **WB**, 1152 logatome files are available for the test. To allow for an adequate test duration, three VCV combinations per speaker (with at most twice the same consonant) are provided to every listener in all telephony conditions. This results in 144 unique logatome files per listener. Every listener has to recognize the correct center consonants of all his logatome test files in an individually randomized order. For both SNRs, a separate test is conducted, each with 8 predominantly non-expert German listeners having a mean age below 30 years and a normal hearing ability. Hence, all of them

[2]Prior to the preprocessing a decimation from 44.1 kHz to 16 kHz sampling rate is required.
[3]Prior to the preprocessing a decimation from 25 kHz to 16 kHz sampling rate is required.

can be considered as *non-native* (with respect to the English syllable pronunciations) and normal-hearing. In this way, the language dependency is analyzed.

Third Test with Native, *Hearing-Impaired* Listeners at 20 dB **and** 0 dB **SNR**

The setup and task of the third subjective syllable articulation test is exactly based on the first one, apart from the following differences (Bauer et al., 2012).

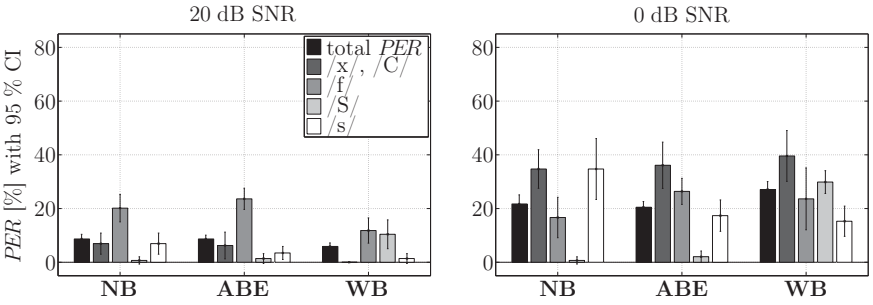
For both SNRs, a separate test is conducted, each with 12 predominantly non-expert German listeners having this time a mean age of 62 years and a monaural hearing impairment. Hence, all of them can be considered as native (with respect to the German syllable pronunciations) and *hearing-impaired*. In this way, the impact of hearing ability is examined. To compensate for their moderate-sloping hearing loss, the listeners are monaurally fitted by high-end, commercially available behind-the-ear (BTE) hearing aids employing the NAL-NL2 formula (Keidser et al., 2011). The test is therefore carried out in the laboratories of `Siemens Audiologische Technik GmbH` (now: `Sivantos GmbH`) by means of an `RME Multiface` external sound card, a high-quality `Genelec` free-field loudspeaker, and a laptop using the `Oldenburg Measurement Applications (OMA)` software from `HörTech gGmbH`.

4.1.2 Experimental Results

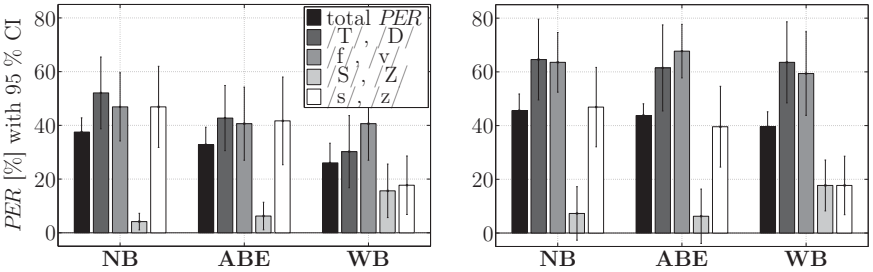
Fig. 4.2 depicts the absolute PER results of the subjective syllable articulation tests with native, normal-hearing listeners (a), *non-native*, normal-hearing listeners (b) as well as native, *hearing-impaired* listeners (c) for 20 dB and 0 dB SNR on the left and right, respectively. Due to the simple test design providing a predefined choice of VCV answers, the resulting PERs only depend on the respective number of phoneme substitutions S_{phn}, i.e., $I_{\text{phn}} = D_{\text{phn}} = 0$ in (3.1). For all telephony conditions, total and phoneme-specific PERs are given, each with a 95 % confidence interval (CI) being defined according to App. B.1. By ignoring the voicing discrimination for the evaluation of the second test, (b) can be compared better to (a) and (c). Following (Bauer et al., 2013), all results are derived without taking into account a correction for guessing (Frary, 1988). Hence, there is a 25 % chance to guess the answers right, which corresponds to a PER of 75 %.

As expected, the PER levels increase consistently in all tests, when reducing the SNR from 20 dB to 0 dB. Since the English VCV pronunciations do not match the native language of the listeners in the second test, the PERs in (b) rise drastically compared with (a) for both SNRs. Due to the hearing impairment of the listeners in the third test, (c) also reveals a remarkably increased PER in relation to (a), particularly for 0 dB SNR.

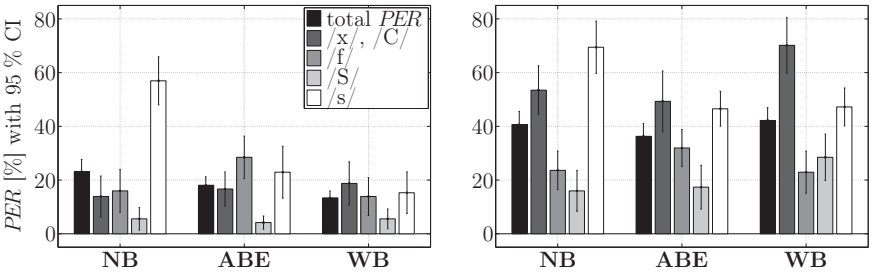
Interestingly, the lowest PERs are not always attained by the **WB** condition. In case of



(a) First test: Native, normal-hearing listeners (Bauer et al., 2013).



(b) Second test: *Non-native*, normal-hearing listeners (Bauer et al., 2010b).



(c) Third test: Native, *hearing-impaired* listeners (Bauer et al., 2012).

Figure 4.2: Total and phoneme-specific PER results of the syllable articulation tests.

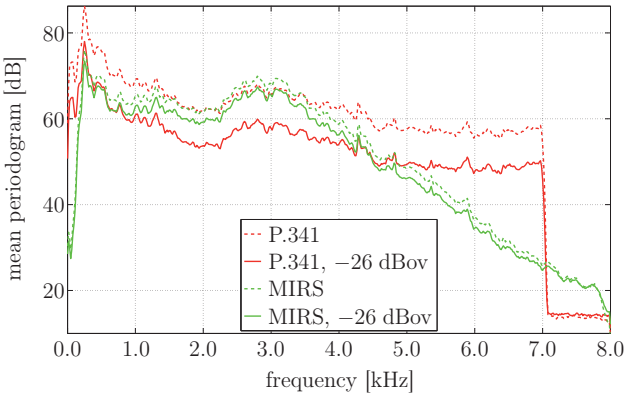
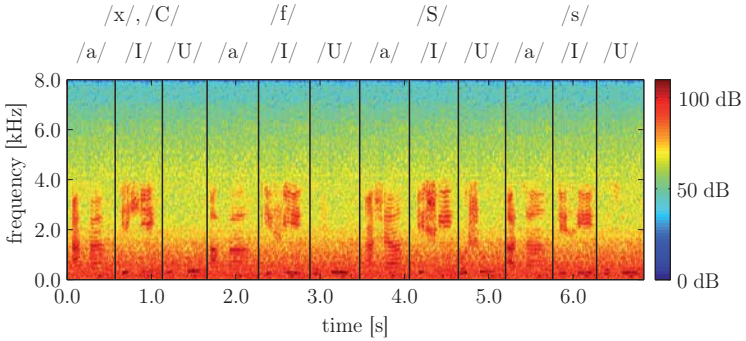


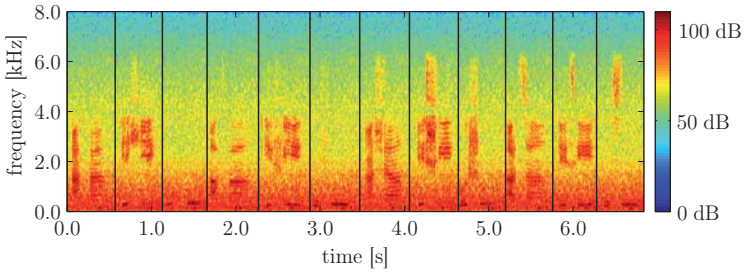
Figure 4.3: Influence of P.341 and MIRS weighting with and without active speech level normalization on the mean periodogram of the German VCV logatomes (Bauer et al., 2013).

the first and third test at 0 dB SNR, **WB** even performs worse than both **ABE** and **NB**. This phenomenon is caused by the (correct) use of the P.341 and MIRS weighting filter masks in Fig. 4.1 to simulate the different sending frequency characteristics of WB- and NB-capable mobile terminals. Their influence on the mean periodogram of the German VCV logatomes is illustrated in Fig. 4.3. Compared with the flat frequency response of the P.341 weighting between 0.05 kHz and 7.0 kHz used for the **WB** condition (ITU-T P.341, 2011), the MIRS weighting applied to the **NB** and **ABE** conditions provokes an attenuation below 1.5 kHz and above 3.5 kHz (ITU-T P.830, 1996, Annex D), as demonstrated by the dashed curves in Fig. 4.3. The low frequencies providing the highest energy contributions are therefore fully preserved only in the P.341-filtered signal. However, this leads to a stronger attenuation, when finally normalizing the active speech levels to −26 dBov for the listening environment preparation in Fig. 4.1. Thus, the normalized, P.341-filtered signal is significantly attenuated in the frequency range of about 1.0 . . . 4.0 kHz relative to the normalized, MIRS-filtered signal, as demonstrated by the solid curves in Fig. 4.3. This scaling effect is particularly relevant for the syllable articulation in the 0 dB SNR case, when the added car noise masks the frequencies below 1.0 kHz to a great extent because of its lowpass characteristics. Please note that the poor performance of the **WB** condition at 0 dB SNR is related to all fricatives except for /s/, which is predominant above 4.0 kHz (Hughes and Halle, 1956; Li and Allen, 2011).

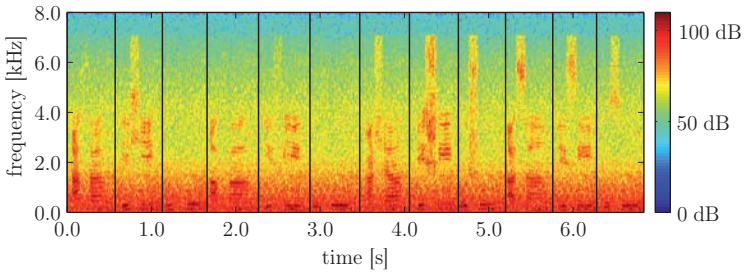
These observations are confirmed by Fig. 4.4. It depicts for all telephony conditions the spectrogram of the concatenated German VCV logatomes pronounced by one exemplary male speaker. Obviously, the aforementioned spectral components between 1.0 kHz and 4.0 kHz are substantially attenuated in case of the **WB** condition (c) compared to the **NB** and



(a) **NB** condition at 0 dB SNR.



(b) **ABE** condition at 0 dB SNR.



(c) **WB** condition at 0 dB SNR.

Figure 4.4: Spectrograms of the German VCV logatomes pronounced by a male speaker.

ABE conditions (a, b). In general, the frequencies above 4.0 kHz are hardly masked by the added car noise as opposed to the frequencies below 1.0 kHz. This explains the good performance of the **WB** condition with respect to fricative /s/ in spite of the scaling effect. Apart from that, the spectrogram of the **ABE** condition (b) is similar to the one of the **WB** condition (c) in the UB frequency range. This indicates that the ABE performs well.

Now let us come back to the absolute PER results depicted in Fig. 4.2. The first test (a) at 20 dB SNR reveals a very low PER level for all telephony conditions. The lowest total PER of 5.90 % is obtained by the **WB** condition, while **NB** and **ABE** both attain 8.68 %. Because of the language match and the normal hearing ability of the listeners, there is not enough potential for ABE to improve the performance of the **NB** condition. When reducing the SNR to 0 dB, the PERs of the fricatives /x/, /C/ and /s/ increase drastically by about 30 percentage points in case of the **NB** condition, whereas the remaining fricatives /f/ and /S/ do not suffer. The **ABE** condition provokes a comparable behavior except for fricative /s/. Its PER rises only by about 15 percentage points. This overcompensates the 10 percentage points higher PER of fricative /f/. Thus, the total PER of the **NB** condition is slightly improved via ABE from 21.70 % to 20.49 %. Due to the aforementioned scaling effect, the highest total PER of 27.08 % results from the **WB** condition.

Because of the language mismatch, the second test (b) offers a huge potential for ABE to improve the **NB** condition at 20 dB SNR, in contrast to the first test (a). All fricatives except for /S/, /Z/ benefit from the **ABE** condition. This results in a total PER reduction from 37.50 % to 32.81 %. The **WB** condition still performs even better with 26.04 %. Mainly responsible for these improvements are the fricatives /T/, /D/ and /s/, /z/. At 0 dB SNR, the PER level of the fricatives /T/, /D/ and /f/, /v/ increases for all telephony conditions to about 60 %. This means that the listeners more or less try to guess them right. In contrast, the fricatives /S/, /Z/ are still recognized well. The PER of the fricatives /s/, /z/ varies most among the telephony conditions. Here, the **ABE** and **WB** conditions outperform the **NB** condition by 7.30 % and 29.17 %, respectively. The total PERs result in 45.57 % for **NB**, 43.75 % for **ABE**, and 39.58 % for **WB**.

As opposed to the first test (a), the hearing impairment of the listeners in the third test (c) also provides a remarkable potential for ABE to improve the **NB** condition at 20 dB SNR. However, fricative /s/ is exclusively responsible for that, whereas the remaining fricatives turn out to be recognized well. Compared to the **NB** condition, **ABE** and **WB** significantly reduce the PER of fricative /s/ by about 30 and 40 percentage points, respectively. In return, ABE degrades the PER of fricative /f/ only by about 10 percentage points. Thus, the total PER of the **NB** condition is improved via ABE from 23.09 % to 18.06 %. A further decrease to 13.37 % is attained by the **WB** condition. At 0 dB SNR, fricative /s/ is tried to be guessed right by the hearing-impaired listeners in the **NB** condition. Once again,

Syllable articulation tests	% total *PER* relative to **NB**			
	20 dB SNR		0 dB SNR	
	ABE	**WB**	**ABE**	**WB**
Native, normal-hearing listeners	±0.00	−32.00	−5.60	+24.80*
Non-native, normal-hearing listeners	−12.50	−30.56	−4.00	−13.14
Native, *hearing-impaired* listeners	−21.80	−42.11	−10.68	+3.85*

Table 4.1: Total PER results relative to the **NB** condition (*: These relative PER degradations are due to the scaling effect as explained in Fig. 4.3 and 4.4).

its PER turns out to be considerably improved based on the **ABE** and **WB** conditions by about 20 percentage points. As the ABE still degrades the PER of fricative /f/ only by about 10 percentage points similar to the 20 dB SNR case, the total PER is reduced by the **ABE** condition from 40.63 % to 36.28 %. The **WB** condition causes the highest total PER of 42.19 % due to the aforementioned scaling effect, like in case of the first test (a) at 0 dB SNR.

Tab. 4.1 finally presents the total PER results of the **ABE** and **WB** conditions relative to the **NB** condition for all subjective syllable articulation tests. Except for the first and third test at 0 dB SNR, the **WB** condition provides the best performance revealing relative PER reductions between 13.14 % and 42.11 %. Obviously, ABE is capable of improving the syllable articulation compared to the **NB** condition in all cases involving low SNR, language mismatch, and hearing impairment. The **ABE** condition thereby attains relative PER improvements ranging from 4.00 % to 21.80 %. These results indicate that ABE is able to significantly narrow the fricative intelligibility gap between NB and WB speech.

4.2 ABE Optimization for Speech Quality Enhancement

Despite the encouraging effect of ABE on syllable articulation, the enhancement of telephone speech quality still poses a big challenge. Telephony users have got accustomed to muffled sounding phone calls over decades. It is therefore difficult to convince them suddenly of a brighter sound. According to Sec. 1.1, a higher speech intelligibility does not necessarily enhance the subjectively perceived quality. Already a few under- and overestimation artifacts, which typically arise from ABE as described in Sec. 1.3, can easily destroy the positive auditive impression. Hence, it is of particular importance that the employed ABE strictly avoids to produce artifacts. From this point of view, several innovations are developed to further optimize the ABE algorithm for the purpose of enhancing telephone speech quality. In the following, an overview of the ABE optimizations is given.

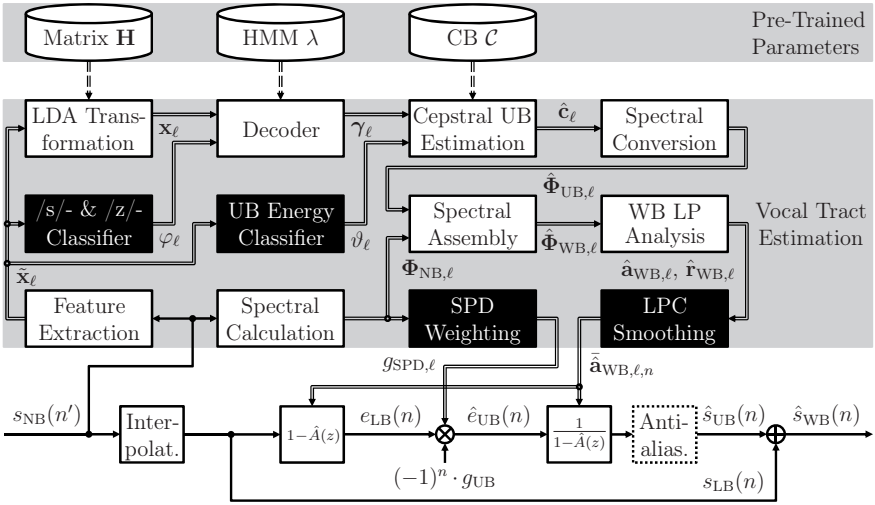


Figure 4.5: Block diagram of optimized ABE processing to enhance telephone speech quality.

4.2.1 Overview

The ABE framework exploiting phonetic *a priori* knowledge in Fig. 2.7 is developed further aiming at a prevention of artifacts. Fig. 4.5 illustrates the block diagram of the optimized ABE processing. Besides the introduction of the blocks filled in black, the serial ABE processing in the lower signal path is replaced by a parallel structure (Vary and Martin, 2006, Fig. 10.5). The coarse ABE postfiltering method being optionally applied at the end of Fig. 2.7 is not required anymore due to novel strategies preventing artifacts in a more sophisticated way. Thus, the upper cut-off frequency reduction yielding a more conservative ABE is eliminated.

Most of the innovations affect the vocal tract estimation part of Fig. 4.5. First of all, two ANN classifier blocks are introduced having access to the composite feature vector $\tilde{\mathbf{x}}_\ell$ before LDA transformation. They shall support the HMM-based estimation process and thus tackle specific problems leading to artifacts. On the one hand, an ANN-based classification of /s/- and /z/-sounds provides the required phoneme class labels φ_ℓ that are not available for ABE processing in the given online telephony application. They can be used for the approved phonetically motivated CB design with two phoneme classes defined in Tab. 3.5. Confusions of the critical fricatives /s/ and /z/ shall be thereby minimized. On the other hand, the energy of the estimated UB spectral envelope is adapted in the cepstral domain via an energy class label ϑ_ℓ. It serves to correct the zeroth LPCC in case of overestimations. Please note that the UB energy estimation also plays an important role in other ABE approaches

(Pulakka, 2013, Sec. 4.5.4). Both ANN classifiers are detailed in Sec. 4.2.2.

Another innovation focuses on the bandwidth extension of speech pauses. Due to their low energy, they can be easily confused with the fricatives /s/ and /z/, when relying on NB speech. These confusions also provoke overestimation artifacts. To tackle this problem, the speech pause extension is suppressed by means of a robust SPD based on the calculated NB power spectrum $\mathbf{\Phi}_{\mathrm{NB},\ell}$. The SF-based residual signal extension is thereby provided with an adaptive weighting factor $g_{\mathrm{SPD},\ell}$. Due to the parallel ABE structure, it only affects the estimated UB residual signal $\hat{e}_{\mathrm{UB}}(n)$. Sec. 4.2.3 presents the SPD-based speech pause extension.

The last block filled in black temporally smooths the estimated WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$ during the first samples after a frame transition. Transients and discontinuities (Välimäki, 1995, Sec. 3.5), which may arise from the time-varying LP analysis and synthesis filtering, shall be thereby reduced (Välimäki and Laakso, 1998). Although they may be hardly audible, they are clearly visible in the spectrogram. The temporal LPC smoothing is actually applied to the corresponding reflection coefficients $\hat{\mathbf{r}}_{\mathrm{WB},\ell}$. Thus, the stability can be verified. A commonly used recursive conversion, which is part of the well-known Levinson-Durbin recursion (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978), subsequently reveals the temporally smoothed, estimated WB LP filter coefficients $\bar{\hat{\mathbf{a}}}_{\mathrm{WB},\ell,n}$. As they depend on sample index n, the LP analysis and synthesis filters therefore adopt them sample-wise. Further details on the temporal LPC smoothing are given in Sec. 4.2.4.

The last innovation is related to the lower signal path and benefits from its parallel ABE structure. As opposed to Fig. 2.7, the optional anti-aliasing filter is not required in Fig. 4.5 for residual signal extension purposes, due to the use of SF. It may be used instead for an IIR highpass filtering of the distorted LB spectrum, after the estimated UB speech $\hat{s}_{\mathrm{UB}}(n)$ has been synthesized. Sec. 4.2.5 demonstrates by means of spectrograms where these alias distortions come from and why such an anti-aliasing therefore makes sense.

4.2.2 ANN-Based HMM Support

According to Sec. 1.3, the most severe artifact typically arising from ABE approaches is the underestimation of /s/ and /z/. It provokes an annoying lisping sound and originates either from a spectral reconstruction using sub-optimally trained CB representatives or from a wrong classification given by *false /s/- and /z/-rejections* (Bauer et al., 2008).

The phonetically motivated CB design with two phoneme classes defined in Tab. 3.5 already tackles the first problem of the sub-optimal spectral reconstruction. It exclusively discriminates between the fricatives /s/, /z/ and all remaining phonemes. For this purpose,

phonetic *a priori* knowledge in terms of the phoneme class labels φ_ℓ is provided to the ABE training, as explained in Sec. 2.1. However, this phonetic support is not available for ABE processing in the given online telephony application. Since there is not sufficient mutual information between the LB and UB frequency range with respect to a one-to-one mapping (Nilsson et al., 2002; Agiomyrgiannakis and Stylianou, 2004), particularly /s/- and /z/-sounds may be misclassified. These misclassifications result in the second problem mentioned above.

This problem can be addressed by detecting the critical fricatives /s/ and /z/ in real time. For this purpose, numerous pattern recognition approaches basically come into consideration (Reif et al., 2014; Fukunaga, 1990). The most promising ones already need to be provided during their training process with the 'ground truth' information to allow for a supervised learning. Amongst others, support vector machines (SVMs) are widely used (Smola and Schölkopf, 2004), however, the number of required support vectors is not determinable prior to the training[4]. In contrast, the computational complexity of ANNs can be already determined during the design phase (Hagan et al., 1996). Despite a relatively complex training, the decoding process of an ANN is in general less complex than an HMM decoder, when assuming comparable topologies. Similar to HMMs, ANNs are commonly applied to phoneme recognition, as exemplarily demonstrated by the TIMIT task (Hinton et al., 2012; Lopes and Perdigão, 2011, Tab. 1, resp. 9). To obtain the required phoneme class labels φ_ℓ in support of the HMM-based estimation process following Sec. 2.2, this work therefore adopts an efficient feed-forward ANN classifier (Bauer et al., 2014a, Sec. 3.2). Please note that ANNs are already successfully utilized in the context of ABE for high-band frequency estimation, too (Pulakka and Alku, 2011; Kontio et al., 2007; Iser et al., 2008).

Due to the reduction of /s/- and /z/-underestimations, however, a contrary artifact is stimulated. Some phonemes – particularly the fricatives /f/, /v/, /S/, /Z/, /x/, /C/, /T/, and /D/, plosives as well as speech pauses – are thereby overestimated. These overestimations provoke undesired hissing and over-voicing artifacts according to Sec. 1.3.

To tackle this problem, Nilsson and Kleijn (2001) augmented their GMM-based ABE algorithm with an asymmetric cost function penalizing underestimations less than overestimations. Alternatively, Pulakka and Alku (2011) used for their ANN-based ABE training an asymmetric error measure. Instead of that, this work employs a second feed-forward ANN classifier to identify high UB energy by means of an energy class label ϑ_ℓ. In case of an overestimation, it serves to adapt the energy of the estimated UB spectral envelope in the cepstral domain (Bauer et al., 2014a, Sec. 3.3). As overestimations basically imply a higher energy in the UB than LB frequency range, they are related to the zeroth LPCC in (2.17).

[4]The computational complexity estimated by some informal SVM experiments on /s/- and /z/-classification was found to be by far too high for the given online ABE application.

Hence, the adaptive UB energy correction is done via $\hat{c}_\ell(0)$.

Imbalanced Two-Class Problems

Both classification problems involve two classes. On the one hand, the first two-class problem distinguishes the critical fricatives /s/ and /z/ denoted by a phoneme class label of $\varphi_\ell = 0$ from all remaining phonemes characterized by $\varphi_\ell = 1$. On the other hand, the second two-class problem differentiates between high UB energy indicated by an energy class label of $\vartheta_\ell = 1$ and low UB energy represented by $\vartheta_\ell = 0$.

Furthermore, the phoneme and energy classes are imbalanced. While the critical fricatives /s/ and /z/ belong to the minority class with $\varphi_\ell = 0$, all remaining phonemes are part of the majority class with $\varphi_\ell = 1$. In the phonetically balanced TIMIT corpus only about 8 % of all phoneme labels represent /s/ and /z/. Correspondingly, the cases of high UB energy expressed by the energy class with $\vartheta_\ell = 1$ are outnumbered compared to the occurrences of $\vartheta_\ell = 0$ implicating less energy in the UB than the LB spectrum.

Randomized sampling methods on the data level can be applied to reduce these class imbalances (He and Garcia, 2009, Sec. 3.1). Thus, balanced data distributions are obtained by either oversampling the minority class or undersampling the majority class. However, this may provoke an overfitting of the oversampled data or information loss of the undersampled data (He and Garcia, 2009, Sec. 3.1.1). Please note that informed sampling methods are used to counteract these effects. For instance, an informed undersampling can be realized by combining several randomly undersampled majority class subsets with the minority class to develop partial classifiers that are finally aggregated (He and Garcia, 2009, Sec. 3.1.2). This procedure underlies the principle of ensemble learning (Opitz and Maclin, 1999) forming the basis of the popular ensemble methods boosting (Schapire, 1990) and bootstrap aggregating (bagging) (Breiman, 1996). Other state-of-the-art methods involving cost-sensitive, kernel-based, active and one-class learning (He and Garcia, 2009, Sec. 3.2–3.4) handle the class imbalance problem on the algorithmic level. However, all of these specialized pattern recognition approaches widely used in the field of machine learning have not been utilized within the scope of this work. As opposed to those classification problems having an absolute class imbalance, the present speech processing application just reveals relative class imbalances, i.e., there is sufficient training speech data available to reasonably model the minority classes (He and Garcia, 2009, Sec. 2).

In case of imbalanced classes, conventional accuracy or error rate measures generally evaluate the classification performance in favor of the majority class. A more differentiated evaluation results from Tab. 4.2 (He and Garcia, 2009, Fig. 9). It illustrates the confusion matrices of the phoneme and energy classification problem on the left and right, respectively.

		Predicted phoneme class		Predicted energy class		
		$\varphi_\ell = 0$	$\varphi_\ell = 1$	$\vartheta_\ell = 1$	$\vartheta_\ell = 0$	
True phoneme class	$\varphi'_\ell = 0$	true acceptance of /s/, /z/	false rejection of /s/, /z/	true acceptance of high UB energy	false rejection of high UB energy	$\vartheta'_\ell = 1$
	$\varphi'_\ell = 1$	false acceptance of /s/, /z/	true rejection of /s/, /z/	false acceptance of high UB energy	true rejection of high UB energy	$\vartheta'_\ell = 0$
						True energy class

Table 4.2: Confusion matrices for performance evaluation of both two-class problems.

The true classes are thereby assigned in vertical direction, while the predicted classes are arranged horizontally. Please note that the true phoneme and energy class labels are denoted by φ'_ℓ and ϑ'_ℓ, respectively. In total, the following four cases arise from each of the two-class problems: True and false acceptance as well as true and false rejection. First of all, the accuracy can be thereby expressed as

$$ACC = \frac{N_{\mathrm{TA}} + N_{\mathrm{TR}}}{N_{\mathrm{TA}} + N_{\mathrm{FR}} + N_{\mathrm{FA}} + N_{\mathrm{TR}}}, \tag{4.1}$$

with N_{TA}, N_{FR}, N_{FA}, and N_{TR} denoting the numbers of cases involving a true acceptance, false rejection, false acceptance, and true rejection, respectively. Based on the accuracy, the error rate results in

$$ER = 1 - ACC = \frac{N_{\mathrm{FA}} + N_{\mathrm{FR}}}{N_{\mathrm{TA}} + N_{\mathrm{FR}} + N_{\mathrm{FA}} + N_{\mathrm{TR}}}. \tag{4.2}$$

When specifically focusing on the cause of each two-class problem, however, false rejection rate (FRR) and false acceptance rate (FAR) measures are more suitable

$$FRR = \frac{N_{\mathrm{FR}}}{N_{\mathrm{FR}} + N_{\mathrm{TA}}}, \tag{4.3}$$

$$FAR = \frac{N_{\mathrm{FA}}}{N_{\mathrm{FA}} + N_{\mathrm{TR}}}. \tag{4.4}$$

On the one hand, the FRR captures well the misclassification problem caused by *false /s/- and /z/-rejections* considering solely the true minority class $\varphi'_\ell = 0$ in Tab. 4.2. On the other hand, a *false acceptance of high UB energy* is responsible for the overestimation problem. It exclusively arises from the true majority class $\vartheta'_\ell = 0$ in Tab. 4.2 and is therefore adequately covered by the FAR.

To treat the respective classes of a two-class problem equally despite its class imbalance, the corresponding FRR and FAR measures can be calibrated by an equal error rate (EER):

$$FRR = FAR = EER. \tag{4.5}$$

Compared with this EER measure, the conventional error rate generally favors the majority class due to its larger amount of data.

Topology and Training of ANN Classifiers

Currently, deep recurrent ANNs are very popular in the field of ASR (Weng et al., 2014; Graves et al., 2013; Hinton et al., 2012). Due to their multiple hidden layers being connected in forward and (at least partially) in backward direction, however, the introduced computational complexity may become too high for practical ABE purposes. Hence, a more efficient feed-forward ANN is designed with one hidden layer (Hagan et al., 1996). The 45-dimensional composite feature vector $\tilde{\mathbf{x}}_\ell$ serves as its input. Furthermore, a single-neuron output layer is used with a saturating linear transfer function to obtain a scalar ANN classifier output in the range of $[0, 1]$.

The remaining parameters have been determined by means of a preliminary full factorial experiment involving the following parameter variations (Bauer et al., 2014a; J. Jones, 2013):

- $N^{(1)} \in \{15, 30, 45, 60\}$ neurons in the hidden layer,
- a saturating linear, log-sigmoid, hyperbolic tangent sigmoid or radial basis transfer function of the hidden layer,
- a resilient or scaled conjugate gradient (SCG) backpropagation training,
- a data ratio between training and validation of 50/50, 70/30 or 90/10.

Based on the resulting 96 ANNs, the best classification performance has been attained by a combination of 45 hidden layer neurons, a saturating linear hidden layer transfer function, an SCG backpropagation training, and a training-to-validation data ratio of 90/10 (Bauer et al., 2014a; J. Jones, 2013). This parametrization is therefore fixed for both ANN classifiers. Their topology is depicted in Fig. 4.6.

As denoted by the dot and dash lines in the hidden layer, the feature elements $\tilde{x}_\ell(\nu)\ \forall\ \nu \in \{1, 2, \ldots, 45\}$ of the input vector $\tilde{\mathbf{x}}_\ell$ are weighted by the elements $w^{(1)}_{\nu,\mu} \in \mathbb{R}\ \forall\ \nu \in \{1, 2, \ldots, 45\}$, $\mu \in \left\{1, 2, \ldots, N^{(1)}\right\}$ of a hidden layer weight matrix $\mathbf{W}^{(1)}$. All products of each hidden layer neuron are then added together with the corresponding element $b^{(1)}_\mu \in \mathbb{R}\ \forall\ \mu \in \left\{1, 2, \ldots, N^{(1)}\right\}$ of a hidden layer bias vector $\mathbf{b}^{(1)}$. Finally, the resulting elements $i^{(1)}_\mu \in \mathbb{R}\ \forall\ \mu \in \left\{1, 2, \ldots, N^{(1)}\right\}$ of the hidden layer net input vector $\mathbf{i}^{(1)}$ are mapped via a saturating linear transfer function $f^{(1)}$ to the elements $o^{(1)}_\mu \in \mathbb{R}\ \forall\ \mu \in \left\{1, 2, \ldots, N^{(1)}\right\}$ of a hidden layer net output vector $\mathbf{o}^{(1)}$.

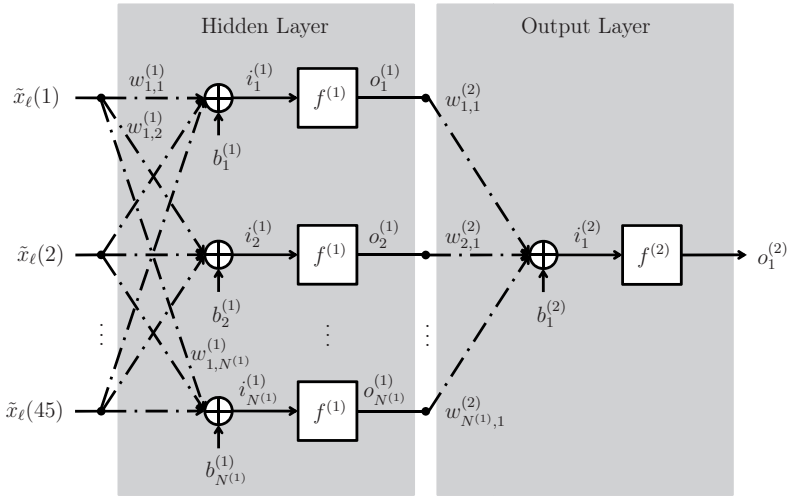


Figure 4.6: Topology of the designed ANN classifiers (here: Hidden layer with $N^{(1)}$ neurons).

The following equations mathematically describe the hidden layer:

$$\begin{bmatrix} i_1^{(1)} \\ i_2^{(1)} \\ \vdots \\ i_{N^{(1)}}^{(1)} \end{bmatrix} = \begin{bmatrix} w_{1,1}^{(1)} & w_{1,2}^{(1)} & \cdots & w_{1,N^{(1)}}^{(1)} \\ w_{2,1}^{(1)} & w_{2,2}^{(1)} & \cdots & w_{2,N^{(1)}}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{45,1}^{(1)} & w_{45,2}^{(1)} & \cdots & w_{45,N^{(1)}}^{(1)} \end{bmatrix}^{\mathrm{T}} \cdot \begin{bmatrix} \tilde{x}_\ell(1) \\ \tilde{x}_\ell(2) \\ \vdots \\ \tilde{x}_\ell(45) \end{bmatrix} + \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_{N^{(1)}}^{(1)} \end{bmatrix}, \tag{4.6}$$

$$o_\mu^{(1)} = f^{(1)}\big(i_\mu^{(1)}\big) = \begin{cases} 0, & \text{if } i_\mu^{(1)} < 0, \\ i_\mu^{(1)}, & \text{if } 0 \leqslant i_\mu^{(1)} \leqslant 1, \\ 1, & \text{if } i_\mu^{(1)} > 1, \end{cases} \quad \forall\, \mu \in \left\{1, 2, \ldots, N^{(1)}\right\}. \tag{4.7}$$

Compared to the hidden layer, the output layer contains only a single neuron. Hence, the elements $o_\mu^{(1)}\ \forall\ \mu \in \left\{1, 2, \ldots, N^{(1)}\right\}$ of the hidden layer net output vector $\mathbf{o}^{(1)}$ are weighted by the elements $w_{\mu,1}^{(2)} \in \mathbb{R}\ \forall\ \mu \in \left\{1, 2, \ldots, N^{(1)}\right\}$ of an output layer weight vector $\mathbf{w}^{(2)}$ and subsequently added together with an output layer bias $b_1^{(2)} \in \mathbb{R}$. The resulting output layer net input $i_1^{(2)} \in \mathbb{R}$ is finally mapped via a saturating linear transfer function $f^{(2)}$ to the output layer net output $o_1^{(2)} \in \mathbb{R}$. A mathematical description of the output layer is given by the

following equations:

$$i_1^{(2)} = \begin{bmatrix} w_{1,1}^{(2)} \\ w_{2,1}^{(2)} \\ \vdots \\ w_{N^{(1)},1}^{(2)} \end{bmatrix}^{\mathrm{T}} \cdot \begin{bmatrix} o_1^{(1)} \\ o_2^{(1)} \\ \vdots \\ o_{N^{(1)}}^{(1)} \end{bmatrix} + b_1^{(2)}, \tag{4.8}$$

$$o_1^{(2)} = f^{(2)}\left(i_1^{(2)}\right) = \begin{cases} 0, & \text{if } i_1^{(2)} < 0, \\ i_1^{(2)}, & \text{if } 0 \leqslant i_1^{(2)} \leqslant 1, \\ 1, & \text{if } i_1^{(2)} > 1. \end{cases} \tag{4.9}$$

For each ANN classifier, the parameters $\left\{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{w}^{(2)}, b_1^{(2)}\right\}$ are separately trained on the predefined training set of the TIMIT corpus (Garofolo et al., 1993) being split into the required training and validation subset[5]. They are thereby adapted to minimize the mean square error (MSE) between the true class labels and the ANN output (J. Jones, 2013, Sec. 2.2.3). ANN training is performed via the neural network toolbox `nntraintool` of `MATLAB R2012a` and automatically terminated as soon as anyone of the following constraints is met:

- 12 successive validations with increasing performance (to prevent overfitting),
- a gradient $\leqslant 10^{-6}$,
- 12,000 training iterations,
- 48 h elapsed time.

ANN evaluation takes place on the predefined TIMIT test set[5]. For both training and test, the employed speech data is preprocessed following Sec. 4.3 to simulate a NB telephony condition. Furthermore, all composite feature vectors $\tilde{\mathbf{x}}_\ell$ are element-wise normalized to a unified range for a faster gradient convergence within training (J. Jones, 2013, Sec. 3.2.2).

ANN for /s/- and /z/-Classification

To allow for a supervised learning, the training of the phonetic ANN classifier is provided with the true phoneme class labels

$$\varphi'_\ell = \begin{cases} 0, & \text{if } \bar{\varphi}_\ell = 0 \wedge c_\ell(0) > 0, \\ 1, & \text{if } \bar{\varphi}_\ell = 1 \vee c_\ell(0) \leqslant 0. \end{cases} \tag{4.10}$$

They rely on the original phoneme class labels $\bar{\varphi}_\ell \in \{0, 1\}$ deriving from the manual phonetic transcriptions of the TIMIT corpus (Garofolo et al., 1993). Moderately sharp /s/- and /z/-pronunciations, which do not reveal more energy in the UB than LB frequency range

[5]Please note that the 'SA' sentences to identify the American dialect region are thereby neglected.

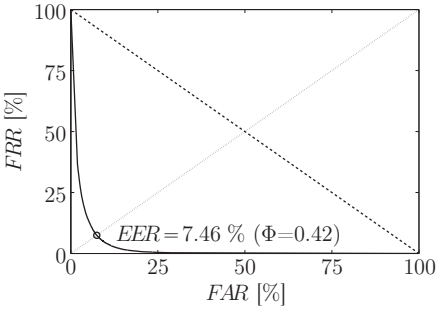


Figure 4.7: DET curve of the ANN-based /s/- and /z/-classifier.

following Sec. 3.2.1, i.e., $c_\ell(0) \leqslant 0$, are assigned to the true majority class $\varphi'_\ell = 1$. Thus, only the sharply pronounced /s/- and /z/-sounds with $c_\ell(0) > 0$ are included in the true minority class $\varphi'_\ell = 0$. Since underestimations are created by *false /s/- and /z/-rejections* according to Tab. 4.2, they only arise from the true minority class. An EER optimization criterion is therefore used for the training of the phonetic ANN classifier to put emphasis on the FRR measure. For this purpose, the class imbalance within the training and validation subset is removed by randomly oversampling the composite feature vectors $\tilde{\mathbf{x}}_\ell$ belonging to $\varphi'_\ell = 0$.

In order to map the real-valued net output $o_1^{(2)} \in [0,\, 1]$ of the ANN output layer to the discrete phoneme class label $\varphi_\ell \in \{0,\, 1\}$, a classification threshold $\Phi \in [0,\, 1]$ is used

$$\varphi_\ell = \begin{cases} 0, & \text{if } o_1^{(2)} < \Phi, \\ 1, & \text{if } o_1^{(2)} \geqslant \Phi. \end{cases} \tag{4.11}$$

The classification performance can be adequately evaluated by a detection error tradeoff (DET) curve (Martin et al., 1997). As an alternative to the commonly used receiver operating characteristic (ROC), it plots the FRR against the FAR. Thus, the value on the main diagonal directly reveals the EER. The closer it is to the origin, the higher is the classification performance. Fig. 4.7 depicts the DET curve of the phonetic ANN classifier, which attains a respectable EER of 7.46 % for $\Phi = 0.42$. Hence, this classification threshold is fixed. A curve close to the dashed line in Fig. 4.7 would have indicated a randomly guessing classifier.

For ABE processing, the acquired phoneme class labels φ_ℓ serve to compute the $N_\mathcal{P} \times N_\mathcal{S}$ phoneme class probability matrix $\mathrm{P}(\varphi_\ell | s_\ell = j)$ in Sec. 2.2.1 with $N_\mathcal{P} = 2$ and $N_\mathcal{S} = 24$. As recommended in Tab. 3.10[6], a phonetic weighting ratio of $r^{(j)} = r = 3 \;\forall\; j \in \mathcal{S}$ is thereby

[6]The FA-based HMM decoder is used for the given online ABE application, since the FBA and VA generally involve too much algorithmic delay.

	ER [%]	*EER* [%]
Without phonetic ANN classifier	6.00	13.17
With phonetic ANN classifier	5.46	10.06

Table 4.3: Performance of HMM-based /s/- and /z/-classifier.

utilized yielding $\varepsilon^{(j)} = \varepsilon = \frac{1}{4} \; \forall \; j \in \mathcal{S}$ according to (2.45). Hence, (2.43) simplifies to

$$\mathrm{P}(\varphi_\ell | s_\ell = j) = \begin{cases} 1 - \varepsilon^{(j)}, & \text{if } (\varphi_\ell = 0 \wedge j \in \mathcal{S}_0) \vee (\varphi_\ell = 1 \wedge j \in \mathcal{S}_1), \\ \varepsilon^{(j)}, & \text{if } (\varphi_\ell = 0 \wedge j \in \mathcal{S}_1) \vee (\varphi_\ell = 1 \wedge j \in \mathcal{S}_0), \end{cases} \quad \forall \; \varphi_\ell \in \mathcal{P}, \; j \in \mathcal{S}. \quad (4.12)$$

Prior to the extensive subjective and instrumental speech quality assessment in Sec. 4.3, the impact of the phonetic ANN classifier on ABE processing is quickly evaluated by investigating the HMM performance in terms of /s/- and /z/-classification. The sum of the *a posteriori* probabilities derived in (2.54) for those states representing /s/ and /z/, i.e., $\sum_{i \in \mathcal{S}_0} \gamma_\ell(i)$, thereby serves as classifier. Tab. 4.3 shows the respective results depending on the support of the phonetic ANN classifier. As the HMM training is performed on imbalanced data with respect to the /s/- and /z/-classification problem, it underlies a conventional error rate optimization criterion. Due to the use of the phonetic ANN classifier, the error rate is reduced from 6.00 % to 5.46 % (for minimum error rate, the classification thresholds of 0.45 and 0.6 have been applied, respectively). Focusing on the EER to equally treat the minority and majority class, the classification performance is improved by the phonetic ANN classifier support from 13.17 % to 10.06 %. The corresponding classification thresholds of 0.02 and 0.01 are relatively low because of the mismatch between optimization criterion and evaluation measure. Nevertheless, the HMM turns out to significantly benefit from the phonetic ANN classifier in terms of /s/- and /z/-classification.

ANN-Based UB Energy Adaptation

The training of the ANN-based UB energy classifier requires for a supervised learning the true energy class labels

$$\vartheta'_\ell = \begin{cases} 1, & \text{if } (c_\ell(0) - 1.0) > 0, \\ 0, & \text{if } (c_\ell(0) - 1.0) \leqslant 0. \end{cases} \quad (4.13)$$

Based on the original zeroth LPCC $c_\ell(0)$ in (2.17), they classify the UB energy into high and low relative to the LB energy. High UB energies turn out to be robustly identified via a conservative thresholding[7] of $(c_\ell(0) - 1.0) > 0$ and are assigned to the minority class $\vartheta'_\ell = 1$. In

[7]Please note that a robust UB energy identification was found to be advantageously conducted by subtracting from $c_\ell(0)$ a constant value of 1.0 before applying a threshold of 0. Thus, only UB energies of $c_\ell(0) > 1.0$ are assumed to be high, which represents a more conservative thresholding than $c_\ell(0) > 0$.

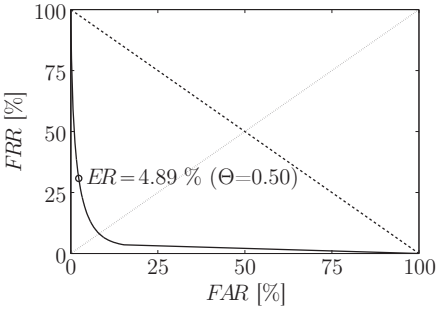


Figure 4.8: DET curve of the ANN-based UB energy classifier.

contrast, $(c_\ell(0)-1.0)\leqslant 0$ indicates low UB energies to be included in the majority class $\vartheta'_\ell = 0$. As overestimation artifacts are provoked by a *false acceptance of high UB energy* according to Tab. 4.2, they originate only from the true majority class and are therefore captured well by the FAR measure. Hence, the class imbalance is preserved, i.e., no randomized sampling methods are applied. The ANN-based UB energy classifier is thus trained under a conventional error rate optimization criterion promoting the majority class.

By using a classification threshold of $\Theta \in [0,\, 1]$, the real-valued net output $o_1^{(2)} \in [0,\, 1]$ of the ANN output layer is mapped to the discrete energy class label $\vartheta_\ell \in \{0,\, 1\}$

$$\vartheta_\ell = \begin{cases} 1, & \text{if } o_1^{(2)} > \Theta, \\ 0, & \text{if } o_1^{(2)} \leqslant \Theta. \end{cases} \tag{4.14}$$

The DET curve of the ANN-based UB energy classifier depicted in Fig. 4.8 points out a respectable classification performance. For $\Theta = 0.5$, it reveals an error rate of 4.89 %. Hence, this classification threshold is fixed. As requested, the FRR is thereby increased in favor of the FAR compared to the EER operation point in Fig. 4.7.

For ABE processing, the acquired energy class labels ϑ_ℓ serve to adaptively correct the energy of the estimated UB cepstral envelope $\hat{\mathbf{c}}_\ell$ in Sec. 2.2.3. For this purpose, the zeroth LPCC is redefined as

$$\hat{c}_\ell(0) := \begin{cases} \hat{c}_\ell(0), & \text{if } \vartheta_\ell = 1, \\ \hat{c}_\ell(0) - 1.0, & \text{if } \vartheta_\ell = 0 \wedge (\hat{c}_\ell(0)-1.0) < 0, \\ (\hat{c}_\ell(0)-1.0) \cdot \left(o_1^{(2)}\right)^\rho, & \text{if } \vartheta_\ell = 0 \wedge (\hat{c}_\ell(0)-1.0) \geqslant 0, \end{cases} \tag{4.15}$$

with $\rho \in \mathbb{R}^+$ being a positive exponent of the real-valued ANN output $o_1^{(2)}$. To simplify matters, the additional use of $o_1^{(2)}$ for UB energy adaptation along with ϑ_ℓ is not specifically illustrated in Fig. 4.5 (see cepstral UB estimation block). According to (4.14), $o_1^{(2)} \in [0,\, \Theta]$ holds for $\vartheta_\ell = 0$ with $\Theta = 0.5$.

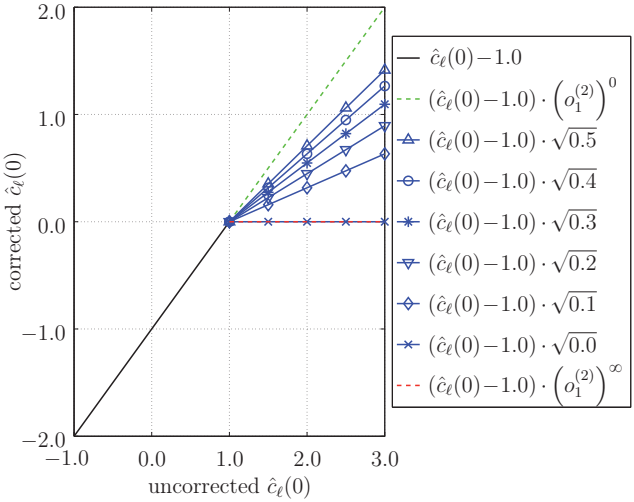


Figure 4.9: ANN-based adaptive UB energy correction for $\vartheta_\ell = 0$, according to (4.15).

Fig. 4.9 aims at visualizing the redefinition of $\hat{c}_\ell(0)$ in (4.15) for $\vartheta_\ell = 0$. In case of $\vartheta_\ell = 1$, the true UB energy is assumed to be high. An overestimation is therefore not expected. Hence, no adaptation of $\hat{c}_\ell(0)$ is undertaken. When assuming the true UB energy to be low, which applies to $\vartheta_\ell = 0$, an overestimation may potentially occur. Hence, an adaptive correction takes place depending on the estimated UB energy.

On the one hand, the estimated UB energy characterized by $(\hat{c}_\ell(0) - 1.0) < 0$ is low, like the expected one. In this case, $\hat{c}_\ell(0)$ is just slightly diminished by a constant subtrahend of 1.0 yielding the black solid line in Fig. 4.9.

On the other hand, $(\hat{c}_\ell(0) - 1.0) \geqslant 0$ indicates that a high UB energy is estimated. The resulting mismatch between the expected low and estimated high UB energy represents a false acceptance according to Tab. 4.2. Thus, an overestimation is identified. In order to reduce it, the real-valued ANN output $o_1^{(2)} \in [0,\, \Theta = 0.5]$ is involved. It is assumed to specify the reliability of the classification $\vartheta_\ell = 0$ in (4.14). Based on this assumption, a high confidence is expressed by values of $o_1^{(2)}$ close to zero, whereas the highest level of uncertainty is reached at $o_1^{(2)} = \Theta = 0.5$. An overestimation is therefore most probable in case of $o_1^{(2)} = 0$. Hence, $o_1^{(2)}$ is predestined to serve as an adaptive attenuation weight in addition to the constant subtraction $\hat{c}_\ell(0) - 1.0$ in (4.15). The blue solid lines in Fig. 4.9 exemplarily show UB energy adaptations for different ANN outputs $o_1^{(2)} \in \{0.0,\, 0.1,\, 0.2,\, 0.3,\, 0.4,\, 0.5\}$ and $\rho = 0.5$. Obviously, the adaptive attenuation rises with decreasing $o_1^{(2)}$, i.e., when overestimations are more likely. A further adjustment can be done by means of the positive exponent ρ within

the range between the red and green dashed lines. A value of $\rho = 0.5$ resulting in $\sqrt{o_1^{(2)}}$ turned out to be adequate.

Prior to the extensive subjective and instrumental speech quality assessment that will be presented in Sec. 4.3, the ability of the ANN-based UB energy adaptation for reducing overestimation artifacts could be verified by several informal listening tests.

4.2.3 SPD-Based Speech Pause Extension

In general, speech pauses are characterized by a low energy. According to Sec. 1.3, ABE approaches therefore tend to confuse them with the critical fricatives /s/ and /z/, when relying on the NB speech signal. Overestimations in terms of undesired hissing artifacts are the consequence. If the ANN-based UB energy classifier described in Sec. 4.2.2 fails, an enhanced extension of speech pauses can be still achieved by means of a robust SPD (Bauer et al., 2014a). According to Fig. 4.5, the speech pause extension is thereby suppressed in the residual domain by an adaptive weighting factor

$$g_{\mathrm{SPD},\ell} = \begin{cases} 0, & \text{if } \mathit{SPD}_\ell = 1, \\ 1, & \text{if } \mathit{SPD}_\ell = 0, \end{cases} \tag{4.16}$$

with $\mathit{SPD}_\ell \in \{0, 1\}$ denoting the hard SPD decision. Obviously, the value of $g_{\mathrm{SPD},\ell}$ is zero in case of a detected speech pause (i.e., $\mathit{SPD}_\ell = 1$) and one for speech presence (i.e., $\mathit{SPD}_\ell = 0$).

Robust SPD

The employed SPD (Fodor and Fingscheidt, 2012, Sec. 4) relies on the frequency-dependent, three-state voice activity detection (VAD)[8] versions (Suhadi, 2012, Sec. 2.2.1) and (Setiawan, 2009, Sec. 4.1.1) for noise power spectral density (PSD) estimation. It detects a speech pause, if both the frame counter ℓ_{SPD}, which specifies the number of frames since the last speech presence has finished, is larger than $L_{\mathrm{SPD}} = 10$ and the smoothed energy $\bar{E}_{\mathrm{SPD},\ell}$ of the current frame does not exceed an SPD threshold $\Theta_{\mathrm{SPD},\ell}$

$$\mathit{SPD}_\ell = \begin{cases} 1, & \text{if } \ell_{\mathrm{SPD}} > L_{\mathrm{SPD}} \wedge \bar{E}_{\mathrm{SPD},\ell} \leqslant \Theta_{\mathrm{SPD},\ell}, \\ 0, & \text{else.} \end{cases} \tag{4.17}$$

The smoothed frame energy is derived from the averaged, pre-emphasized NB power spectrum $\bar{\Phi}_{\mathrm{NB},\ell}(k')$ by the sum over $k' = k'_{\mathrm{low}}, k'_{\mathrm{low}}+1, \ldots, k'_{\mathrm{high}}-1$

$$\bar{E}_{\mathrm{SPD},\ell} = \sum_{k'=k'_{\mathrm{low}}}^{k'_{\mathrm{high}}-1} \bar{\Phi}_{\mathrm{NB},\ell}(k'), \tag{4.18}$$

[8] In contrast to the SPD, a VAD primarily aims at detecting speech activity.

with k'_{low} and k'_{high} denoting the frequency bins $\frac{0.1\text{ kHz}}{f'_s/2} \cdot N'_w/2 = 2$ and $\frac{4.0\text{ kHz}}{f'_s/2} \cdot N'_w/2 = 80$, respectively. Thus, noise components being usually high at low frequencies are not taken into account until 0.1 kHz. A first-order IIR recursion with a smoothing factor of $\alpha_{\text{SPD}} = 0.5$ is utilized to smooth the NB power spectrum $\Phi_{\text{NB},\ell}(k')$ for $k' = 0, 1, \ldots, N'_w - 1$

$$\bar{\Phi}_{\text{NB},\ell}(k') = \alpha_{\text{SPD}} \cdot \bar{\Phi}_{\text{NB},\ell-1}(k') + (1-\alpha_{\text{SPD}}) \cdot \Phi_{\text{NB},\ell}(k') \left|H_{\text{HP}}(k')\right|^2. \tag{4.19}$$

To counteract the lowpass characteristics of the NB speech signal, the higher frequency components are slightly accentuated by pre-emphasizing $\Phi_{\text{NB},\ell}(k')$ via the frequency response $H_{\text{HP}}(k')$ of a first-order FIR highpass filter with the coefficients $\mathbf{b}_{\text{HP}} = [1, -0.8]^{\text{T}}$. The averaged, pre-emphasized NB power spectrum is initialized by $\bar{\Phi}_{\text{NB},0}(k') = \Phi_{\text{NB},0}(k') \left|H_{\text{HP}}(k')\right|^2$ $\forall\ k' = 0, 1, \ldots, N'_w - 1$.

The SPD threshold is computed frame by frame using an adaptive noise floor signal $E_{\text{flr},\ell}$ as well as the constant parameters $M_{\text{flr}} = 9.6$ and $A_{\text{flr}} = 2.5$

$$\Theta_{\text{SPD},\ell} = M_{\text{flr}} \cdot E_{\text{flr},\ell} + A_{\text{flr}}. \tag{4.20}$$

Only in case of an enduring speech absence, which can be assumed when ℓ_{SPD} is larger than L_{SPD} as well as $\bar{E}_{\text{SPD},\ell}$ does not exceed twice a control signal $V_{\text{ctr},\ell}$, the noise floor signal will be updated by $\bar{E}_{\text{SPD},\ell}$

$$E_{\text{flr},\ell} = \begin{cases} \beta_{\text{SPD}} \cdot E_{\text{flr},\ell-1} + (1-\beta_{\text{SPD}}) \cdot \bar{E}_{\text{SPD},\ell}, & \text{if } \ell_{\text{SPD}} > L_{\text{SPD}} \wedge \bar{E}_{\text{SPD},\ell} \leqslant 2V_{\text{ctr},\ell}, \\ E_{\text{flr},\ell-1}, & \text{else.} \end{cases} \tag{4.21}$$

This update is conducted via a first-order IIR recursion with a smoothing factor of β_{SPD}

$$\beta_{\text{SPD}} = \begin{cases} 0.5, & \text{if } \bar{E}_{\text{SPD},\ell} \leqslant E_{\text{flr},\ell-1}, \\ 0.9375, & \text{else.} \end{cases} \tag{4.22}$$

If $E_{\text{flr},\ell-1}$ is smaller than $\bar{E}_{\text{SPD},\ell}$, the smoothing factor is close to one and $E_{\text{flr},\ell}$ therefore highly depends on $E_{\text{flr},\ell-1}$. Otherwise, the updated noise floor signal results from the average of $E_{\text{flr},\ell-1}$ and $\bar{E}_{\text{SPD},\ell}$. The control signal tracks the valleys of the smoothed frame energy and is steadily raised by the constant factor $M_{\text{ctr}} = 1.025$ during speech activity

$$V_{\text{ctr},\ell} = \begin{cases} \bar{E}_{\text{SPD},\ell}, & \text{if } \bar{E}_{\text{SPD},\ell} < V_{\text{ctr},\ell-1}, \\ M_{\text{ctr}} \cdot V_{\text{ctr},\ell-1}, & \text{if } \bar{E}_{\text{SPD},\ell} > 2V_{\text{ctr},\ell-1}, \\ V_{\text{ctr},\ell-1}, & \text{else.} \end{cases} \tag{4.23}$$

Thus, a suddenly beginning and then stationary noise signal will not be detected as speech presence after a transition time. For initialization purposes, the control and floor signals are set to $V_{\text{ctr},0} = \bar{E}_{\text{SPD},0}$ and $E_{\text{flr},0} = 2\bar{E}_{\text{SPD},0}$, respectively.

After having implemented and parameter-optimized the SPD, it has been extensively tested to ensure noise robustness. The tests involved stationary and instationary noise types for different SNRs in combination with two commonly used noise reduction algorithms[9].

Residual-Domain SPD Weighting

The adaptive weighting factor $g_{\mathrm{SPD},\ell}$ derived via (4.16) from the SPD decision SPD_ℓ in (4.17) is applied to the SF-based residual signal extension, as depicted in Fig. 4.5. Thus, the sample-wise estimation of the UB residual signal in (2.84) is modified as follows

$$\hat{e}_{\mathrm{UB}}(n) = e_{\mathrm{LB}}(n) \cdot (-1)^n \cdot g_{\mathrm{UB}} \cdot g_{\mathrm{SPD},\ell}. \tag{4.24}$$

To suppress the extension of speech pauses, $g_{\mathrm{SPD},\ell}$ adaptively weights the UB residual signal jointly with the static attenuation factor g_{UB}. Please note that $g_{\mathrm{SPD},\ell}$ is thereby renewed every frame $\ell = 0, 1, \ldots, L-1$ at the sample indices $n = \ell \cdot N_{\mathrm{s}}$.

4.2.4 Temporal LPC Smoothing

According to (Välimäki, 1995; Schnell and Lacroix, 2008), transients and discontinuities may arise from the time-variant LP analysis and synthesis filtering. In order to reduce these switching effects, Sec. 2.2.4 performs a moderate cepstral interframe smoothing (2.69)–(2.70) and Sec. 2.2.8 recommends the use of a non-transposed LP analysis and synthesis filter structure in (2.79)–(2.80). However, the introduced problems are thereby not completely solved yet.

Inspired by (Välimäki and Laakso, 1998), the estimated WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$ are therefore temporally smoothed during the first N_{t} samples after a frame transition. For the purpose of stability verification following (2.76), this temporal smoothing is actually conducted based on the corresponding reflection coefficients $\hat{\mathbf{r}}_{\mathrm{WB},\ell-1}$ and $\hat{\mathbf{r}}_{\mathrm{WB},\ell}$ of the previous and current frame, respectively:

$$\bar{\hat{\mathbf{r}}}_{\mathrm{WB},\ell,n} = \begin{cases} \left(1 - \frac{n-\ell N_{\mathrm{s}}}{N_{\mathrm{t}}}\right) \cdot \hat{\mathbf{r}}_{\mathrm{WB},\ell-1} + \frac{n-\ell N_{\mathrm{s}}}{N_{\mathrm{t}}} \cdot \hat{\mathbf{r}}_{\mathrm{WB},\ell}, & \text{if } \ell N_{\mathrm{s}} \leqslant n \leqslant \ell N_{\mathrm{s}} + N_{\mathrm{t}} - 1, \\ \hat{\mathbf{r}}_{\mathrm{WB},\ell}, & \text{if } \ell N_{\mathrm{s}} + N_{\mathrm{t}} \leqslant n \leqslant (\ell+1) N_{\mathrm{s}} - 1. \end{cases} \tag{4.25}$$

The resulting reflection coefficients $\bar{\hat{\mathbf{r}}}_{\mathrm{WB},\ell,n}$ are thereby averaged sample by sample to smoothly switch over from $\hat{\mathbf{r}}_{\mathrm{WB},\ell-1}$ to $\hat{\mathbf{r}}_{\mathrm{WB},\ell}$. Hence, they depend on sample index n. After N_{t} samples, the smooth coefficient switch is practically finished. It turns out that $N_{\mathrm{t}} = N_{\mathrm{LP}^{(\mathrm{WB})}}$ samples

[9] Both of them derive the noise PSD via minimum statistics (Martin, 2001) and perform a decision-directed *a priori* SNR estimation (Ephraim and Malah, 1984). The weighting rules are based on a Wiener filter (Scalart and Filho, 1996) and super-Gaussian joint MAP estimator (Lotter and Vary, 2005), respectively.

provides a good trade-off between smoothness and accurateness of $\bar{\tilde{\mathbf{r}}}_{\mathrm{WB},\ell,n}$. The temporally smoothed reflection coefficients for $\ell=0$ are initialized by $\bar{\tilde{\mathbf{r}}}_{\mathrm{WB},0,n}=\hat{\mathbf{r}}_{\mathrm{WB},0}\ \forall\ n=0,1,\ldots,N_{\mathrm{s}}{-}1$.

After the stability of $\bar{\tilde{\mathbf{r}}}_{\mathrm{WB},\ell,n}\ \forall\ n=\ell N_{\mathrm{s}},\ell N_{\mathrm{s}}{+}1,\ldots,(\ell{+}1)N_{\mathrm{s}}{-}1$ has been verified successfully[10], a commonly used recursive conversion into the corresponding temporally smoothed estimated WB LP filter coefficients $\bar{\tilde{\mathbf{a}}}_{\mathrm{WB},\ell,n}$ takes place, being part of the well-known Levinson-Durbin recursion (Makhoul, 1975; Markel and Gray, 1976; Rabiner and Schafer, 1978). In contrast to (2.79)–(2.80), the resulting filter coefficients $\bar{\tilde{\mathbf{a}}}_{\mathrm{WB},\ell,n}$ are adapted for LP analysis and synthesis filtering at every sample n because of their dependency on sample index n

$$e_{\mathrm{LB}}(n)=s_{\mathrm{LB}}(n)-\sum_{\nu=1}^{N_{\mathrm{LP(WB)}}}\bar{\tilde{a}}_{\mathrm{WB},\ell,n}(\nu)\cdot s_{\mathrm{LB}}(n-\nu), \tag{4.26}$$

$$\hat{s}_{\mathrm{UB}}(n)=\hat{e}_{\mathrm{UB}}(n)+\sum_{\nu=1}^{N_{\mathrm{LP(WB)}}}\bar{\tilde{a}}_{\mathrm{WB},\ell,n}(\nu)\cdot \hat{s}_{\mathrm{UB}}(n-\nu). \tag{4.27}$$

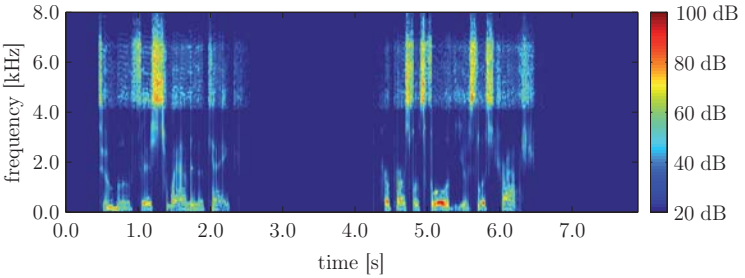
Due to the parallel ABE structure in Fig. 4.5, the LP synthesis filter is only applied in (4.27) to the estimated UB residual signal $\hat{e}_{\mathrm{UB}}(n)$ and thus computes the estimated UB speech $\hat{s}_{\mathrm{UB}}(n)$, while the LB speech $s_{\mathrm{LB}}(n)$ is bypassed.

Fig. 4.10 (a) and (b) depict the UB speech spectrograms of an English utterance without and with temporal LPC smoothing, respectively. When comparing both of them with the LB speech spectrogram in (c), a reliable ABE suppression during speech absence via the SPD-based speech pause extension in Sec. 4.2.3 becomes obvious. Please note further that the switching effects caused by the time-variant LP analysis and synthesis filtering are clearly visible in the LB frequency range of (a). Fortunately, they are significantly reduced by the temporal LPC smoothing in (b). The remaining alias distortions below about 1.0 kHz do not originate from the time-varying filter coefficients and are addressed in the next section.

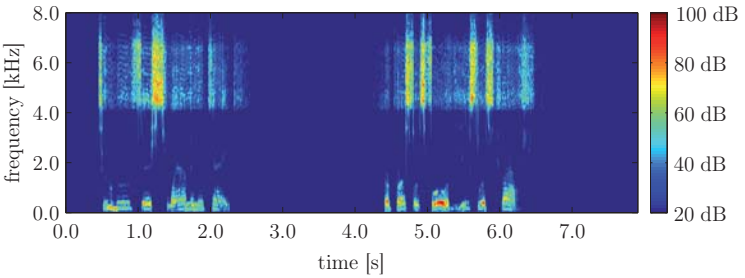
4.2.5 Anti-Aliasing of LB Spectrum

The cause of the remaining low-frequency alias distortions in Fig. 4.10b is based on the spectral characteristic of the employed WB speech for ABE training. It reveals both an increase of energy towards low frequencies (spectral tilt), which is typical for speech signals, and a strong decrease of energy for frequencies above 7.0 kHz (ITU-T P.341, 2011). The latter is modeled by the UB cepstral envelopes for CB training in Sec. 2.1.2 and therefore influences the shape of the UB spectral envelope estimated within ABE processing. When assembling the estimated WB power spectrum in Sec. 2.2.5, the low-frequency energy gain is merged with the high-frequency energy drop. Hence, the estimated WB LP filter coefficients imply both. First of all, the LP analysis filter applies to $s_{\mathrm{LB}}(n)$ the spectrally inverted

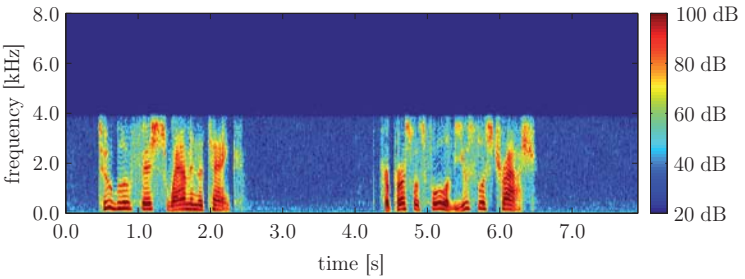
[10]Please note that the temporal LPC smoothing did never produce instabilities within this work.



(a) $\hat{s}_{\mathrm{UB}}(n)$ without temporal LPC smoothing.



(b) $\hat{s}_{\mathrm{UB}}(n)$ with temporal LPC smoothing.



(c) $s_{\mathrm{LB}}(n)$.

Figure 4.10: UB and LB speech spectrograms of the following utterance (NTT-AT, 1994): "Two blue fish swam in a tank, her purse was full of useless trash."

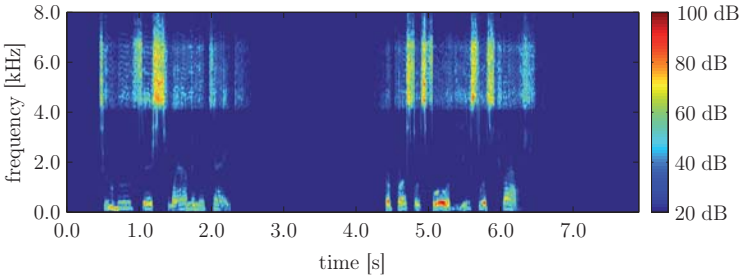
envelope, i.e., the low frequencies are attenuated and the high frequencies amplified. As the UB spectrum in $s_{\mathrm{LB}}(n)$ does not have an infinitely high stopband attenuation, the weak spectral components above 7.0 kHz are thereby emphasized in $e_{\mathrm{LB}}(n)$. Due to the SF-based residual signal extension in Fig. 4.5, they are spectrally mirrored to the frequency range below 1.0 kHz in $\hat{e}_{\mathrm{UB}}(n)$. Finally, the LP synthesis filter amplifies these alias components again, but this time by means of the low-frequency energy gain. Thus, the LB alias distortions in $\hat{s}_{\mathrm{UB}}(n)$ originate from a twofold amplification combined with a spectral mirroring.

The introduced distortions can be reduced by anti-aliasing the LB spectrum in $\hat{s}_{\mathrm{UB}}(n)$. The optional anti-aliasing filter in Fig. 2.7, which is not required by the SF-based residual signal extension, is therefore applied to $\hat{s}_{\mathrm{UB}}(n)$ in Fig. 4.5. For this purpose, it is redesigned as a fifth-order, elliptic IIR highpass filter with a stopband attenuation of 100 dB, a passband ripple of 0.1 dB, and a transition range of $1.0 \ldots 4.0$ kHz (Oppenheim and Schafer, 1989, App. B.3). This filter design is of low computational complexity and algorithmic delay. Please note that the LB spectrum anti-aliasing benefits from a decomposition of $\hat{s}_{\mathrm{UB}}(n)$ and $s_{\mathrm{LB}}(n)$ due to the parallel ABE structure. If such an anti-aliasing was conducted after LP synthesis filtering in the serial ABE structure of Fig. 2.7, the transparency of the LB spectrum in $\hat{s}_{\mathrm{WB}}(n)$ would be completely destroyed.

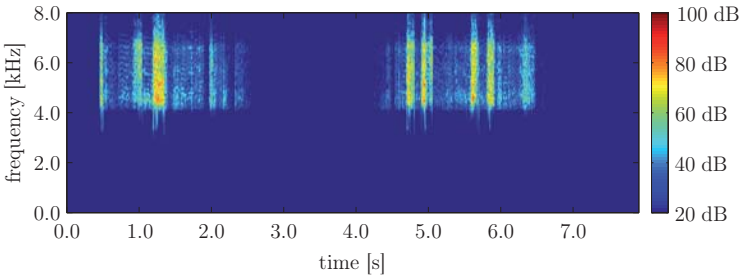
For the purpose of comparison, the UB speech spectrogram with temporal LPC smoothing of Fig. 4.10b is illustrated again in Fig. 4.11a. Obviously, its alias distortions below about 1.0 kHz are completely removed by the LB spectrum anti-aliasing in Fig. 4.11b. According to Fig. 4.5, the undistorted $\hat{s}_{\mathrm{UB}}(n)$ is therefore superposed with $s_{\mathrm{LB}}(n)$ yielding the ABE speech spectrogram in Fig. 4.11c. It implies all ABE optimizations proposed in Sec. 4.2 aiming at the reduction of artifacts. In the following, a subjective and instrumental evaluation is conducted to extensively assess the ABE performance in terms of speech quality.

4.3 Speech Quality Assessment

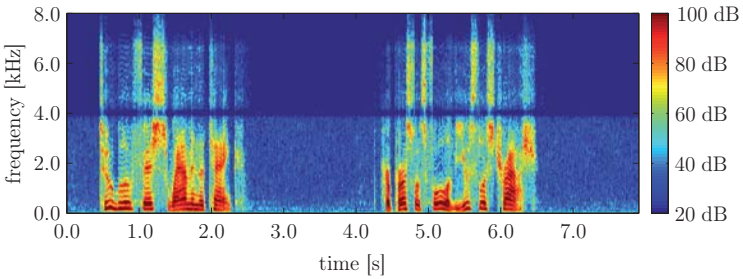
After the ABE has been optimized to reduce different kinds of artifacts, i.e., under- as well as overestimations, filter switching effects, and alias distortions, this section investigates the resultant impact on speech quality. For the purpose of speech quality evaluation, commonly used subjective and instrumental assessment methods are involved. The subjective speech quality perception is assumed to represent the ground truth for the given human-to-human ABE application and needs to be predicted as accurately as possible by the instrumental measurements. Two main objectives are therefore pursued. On the one hand, it shall be clarified, whether ABE is capable of enhancing NB telephone speech quality, apart from the attained speech intelligibility improvement demonstrated in Sec. 4.1. On the other hand, it shall be found out, how to reliably assess ABE systems in practice.



(a) $\hat{s}_{\mathrm{UB}}(n)$ with temporal LPC smoothing but without LB spectrum anti-aliasing.



(b) $\hat{s}_{\mathrm{UB}}(n)$ with both temporal LPC smoothing and LB spectrum anti-aliasing.



(c) $\hat{s}_{\mathrm{WB}}(n)$ with both temporal LPC smoothing and LB spectrum anti-aliasing.

Figure 4.11: UB and ABE speech spectrograms of the following utterance (NTT-AT, 1994): "Two blue fish swam in a tank, her purse was full of useless trash."

4.3.1 Subjective Speech Quality Tests

Following (ITU-T P.800, 1996, Sec. 6), either conversation-opinion or listening-opinion tests come into consideration to subjectively assess ABE systems in terms of speech quality. According to (Pulakka, 2013, Sec. 6.1.4), conversational evaluations of ABE have been only conducted in (Laaksonen et al., 2011; Pulakka et al., 2012a,b,c). Compared to listening-only tests, they better reflect telephone conversations. However, their sensitivity may be lower, since the participating conversational partners do not only have to focus on the ratings but also on the respective conversations. Furthermore, the arrangement of conversation-opinion tests is even more complex (ITU-T P.800, 1996, Annex A). Hence, listening-opinion tests are commonly used instead (Pulakka, 2013, Sec. 6.1.1–6.1.2). Preference and similarity tests are their simplest representatives (Pulakka, 2013, Sec. 6.1.2). ITU-R BS.1534-1 (2003) recommends a more sophisticated multi-stimulus test with hidden reference and anchor (MUSHRA). It principally aims at assessing audio codecs, but has also been sporadically used for ABE evaluation (Pulakka, 2013, Sec. 6.1.1). Its high-resolution continuous quality scale (CQS) requires expert listeners, who are experienced in detecting small impairments but not aware of the test purposes. The majority of listening-only tests for ABE evaluation, however, rely on the much coarser absolute category rating (ACR), degradation category rating (DCR), and comparison category rating (CCR) scales (Pulakka, 2013, Sec. 6.1.1). They are recommended by (ITU-T P.800, 1996, Annex B, D–E) mainly for the purpose of speech codec assessment (ITU-T P.830, 1996). For instance, the standardization process of the WB AMR speech codec included ACR and DCR listening tests in the qualification (pre-selection) phase, while the (final) selection phase consisted of ACR, DCR, and CCR listening tests (3GPP S4/SMG11, 2000; 3GPP S4, 2000; 3GPP SA WG4, 2000).

Inspired by the two-stage speech codec qualification and selection, a methodology for the subjective speech quality assessment of ABE systems is proposed in (Bauer et al., 2014c). It first of all involves the qualification phase in terms of an ACR listening test. Potential ABE candidates that have thereby qualified for the selection phase are finally compared with each other by means of a CCR listening test. As ABE systems do (hopefully) not degrade the speech quality of their input signals in contrast to speech codecs, DCR listening tests are completely excluded. The resulting two-stage subjective ABE evaluation process slightly refined in (Bauer et al., 2014a, Sec. 4) serves as the basis for the remainder of this section.

Preprocessing of Speech Data

The subjective speech quality tests include two female and two male German voices. For each of these speakers, four utterances of about 8 s are provided. This results in a total of 16 speech signals sampled at 16 kHz. They originate from the German part of the NTT-AT

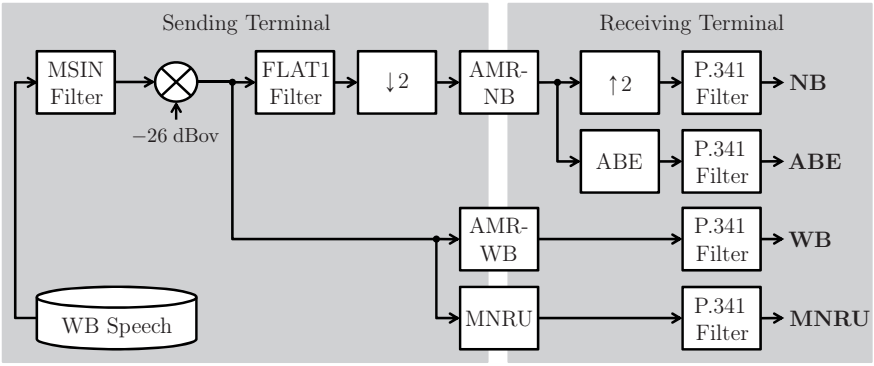


Figure 4.12: Data preprocessing for the subjective speech quality tests simulating realistic NB, ABE-enhanced and WB telephony conditions as well as MNRU-based reference anchors.

multi-lingual speech database for telephonometry (NTT-AT, 1994) and are preprocessed according to Fig. 4.12. Realistic NB, ABE-enhanced and WB telephony conditions are thereby simulated taking into account the transmission characteristics of the sending and receiving terminal including the speech codec. In total, one **NB**, six **ABE** and three **WB** conditions are created. Furthermore, six **MNRU** conditions serve as reference anchors for the ACR test to fully exploit the dynamic range of the 5-point listening-quality scale introduced in Sec. 1.1 (ITU-T P.800, 1996, Annex B). They imply different ratios of speech to modulated noise power adjusted via MNRU (ITU-T P.810, 1996). All algorithmic delay contributions in Fig. 4.12 are exactly compensated for to ensure a temporal synchronization of the telephony conditions.

Following (Pulakka and Alku, 2011), a flat highpass filter with a cut-off frequency of about 0.2 kHz is applied at first to each speech file via MSIN (ITU-T G.191, 2009). It models the input characteristics of both NB and WB mobile terminals fulfilling the P.341 hands-free sending sensitivity/frequency mask (ITU-T P.341, 2011, Fig. 9). Furthermore, the file-based active speech levels are consistently scaled via (ITU-T G.191, 2009) to −26 dBov (ITU-T P.56, 2011). The normalized WB speech files form the basis of all 16 telephony conditions. Thus, the scaling effect arising from Fig. 4.3 due to the preprocessing in Sec. 4.1.1 is prevented.

In the upper part of Fig. 4.12, a flat lowpass filter with a cut-off frequency of about 3.6 kHz completes the modeled sending frequency characteristics of NB-capable mobile terminals, according to the data preparation for the large-vocabulary ASR experiments in Sec. 3.4.1. It is derived from the so-called FLAT1 filter in (ITU-T G.191, 2009). The lowpass-filtered speech signals are decimated to a sampling rate of $f_s' = 8$ kHz. By applying the NB AMR

speech codec at the commonly used bit rate of 12.2 kbps (3GPP TS 26.090, 1999), a NB mobile telephony call is simulated. The resulting speech signals are subject to an interpolation of factor two in parallel with ABE processing. The latter takes into account six ABE candidates:

1a: Original ABE implementation after (Jax, 2002) revealing the best baseline ABE performance in (Bauer et al., 2014c, Fig. 1: *ABE1b*)[11].

2a: Interim ABE implementation used for the syllable articulation tests in Sec. 4.1, i.e., without including any ABE optimization of Sec. 4.2.

2b: Interim ABE implementation further developed from 2a revealing the best overall ABE performance in (Bauer et al., 2014c, Fig. 1: *ABE2C*)[12].

3a: Final ABE implementation further developed from 2a including all ABE optimizations of Sec. 4.2[13] except for the ANN-based UB energy adaptation in Sec. 4.2.2.

3b: Final ABE implementation further developed from 2a including all ABE optimizations of Sec. 4.2[13] except for the ANN-based /s/- and /z/-classification in Sec. 4.2.2.

3c: Final ABE implementation further developed from 2a including all ABE optimizations of Sec. 4.2[13].

Please note that ABE candidate 1a forms the basis of the remaining ABE candidates. It therefore serves as the *reference* ABE implementation for the subsequent ACR and CCR listening tests. Among the *proposed* ABE implementations 2a–3c, the one with the best ACR test performance joins the CCR test as well. Each ABE candidate is trained on all 5578 phonetically rich sentences available for the close-talk channel of the American English speech corpus SpeechDat-Car US (Moreno et al., 2000). This amounts to approximately 4.0 h of NB and WB speech data being preprocessed largely following Fig. 4.12. The introduced language mismatch between ABE training and processing poses a challenge all ABE candidates have to cope with.

In the lower part of Fig. 4.12, the WB AMR speech codec is applied using three bit rates (3GPP TS 26.190, 2001): 8.85 kbps, 12.65 kbps, and 23.85 kbps. They simulate a WB mobile telephony call with low, sufficient, and high quality, respectively. Apart from that, the reference anchors are prepared via MNRU (ITU-T P.810, 1996) adjusting six ratios of speech to modulated noise power: ∞ dB (i.e., clean), 45 dB, 35 dB, 25 dB, 15 dB, and 5 dB.

Finally, the receiving frequency characteristics of WB-capable mobile terminals is equally modeled for all 16 speech conditions. Similar to (Pulakka and Alku, 2011), it is simulated by a flat bandpass filtering to the frequency range of about $0.2 \ldots 7$ kHz, which fulfills the P.341 hands-free receiving sensitivity/frequency mask (ITU-T P.341, 2011, Fig. 10).

[11] For specific details on implementation and parametrization please refer to (Bauer et al., 2014c, Sec. 2.1).
[12] For specific details on implementation and parametrization please refer to (Bauer et al., 2014c, Sec. 2.2).
[13] According to Sec. 4.2.1, the optional ABE postfiltering of Sec. 2.2.10 is not used here anymore.

ACR Test Setup

At first, an ACR listening test is conducted largely following (ITU-T P.800, 1996, Annex B). It engages eight female as well as sixteen male German students of age under 30 years, and with normal hearing abilities. All 24 predominantly non-expert listeners are compensated for participation by a service charge of 10 € for about 30 min instruction, familiarization, and actual test. They are assigned to two listening panels providing three test sessions each, i.e., always four subjects participate together in a test session. Before each test session, the listeners are briefly instructed according to (ITU-T P.800, 1996, Annex B). One utterance per speaker is spent to individually familiarize the listeners of each test session with the test procedure. Based on all 64 preprocessed speech files deriving from these four utterances, the familiarizations consider each telephony condition only once. The remaining three utterances per speaker result in totally 192 preprocessed speech files and are dedicated to the actual test. They are equally divided into both listening panels allowing for a balanced occurrence of speaker-specific utterances and telephony conditions.

Each test session begins with an instruction and a familiarization phase followed by the actual test. The respective speech files for familiarization and test are concatenated in randomized order. It is assured that two successive speech files do not originate from the same utterance. After each appended file, a silence sequence of 5 s is inserted to provide enough time for rating. The active speech level of the concatenated files is scaled via (ITU-T G.191, 2009) to −26 dBov (ITU-T P.56, 2011) and interpolated to 48 kHz sampling rate. Finally, 48 kHz-sampled car noise scaled via (ITU-T G.191, 2009) to an RMS level of −66 dBov is added to adjust an SNR of 40 dB (ITU-T P.56, 2011). This carefully masks the idle noise arising from the NB AMR speech codec. In total, six concatenated familiarization and actual test files are thereby created. The resulting files are diotically presented in a quiet room via four high-quality `AKG K 271 MK II` headphones using a laptop with an `RME Fireface 400` external sound card and a `SAMSON S-phone` multichannel headphone amplifier. A comfortable sound level can be individually adjusted during the familiarization phase. Each test session lasts about half an hour for instruction, familiarization, and actual test.

ACR Test Results

For all telephony conditions of the ACR listening test, an MOS with 95 % CI following App. B.1 is illustrated in Fig. 4.13. As expected, the MNRU-based reference anchors 11–16 almost fully exploit the dynamic range of the listening-quality scale from 1 (bad) to 5 (excellent). When increasing the ratio of speech to modulated noise power, the MOS value rises consistently. Hence, the highest MOS is attained by the clean reference anchor with

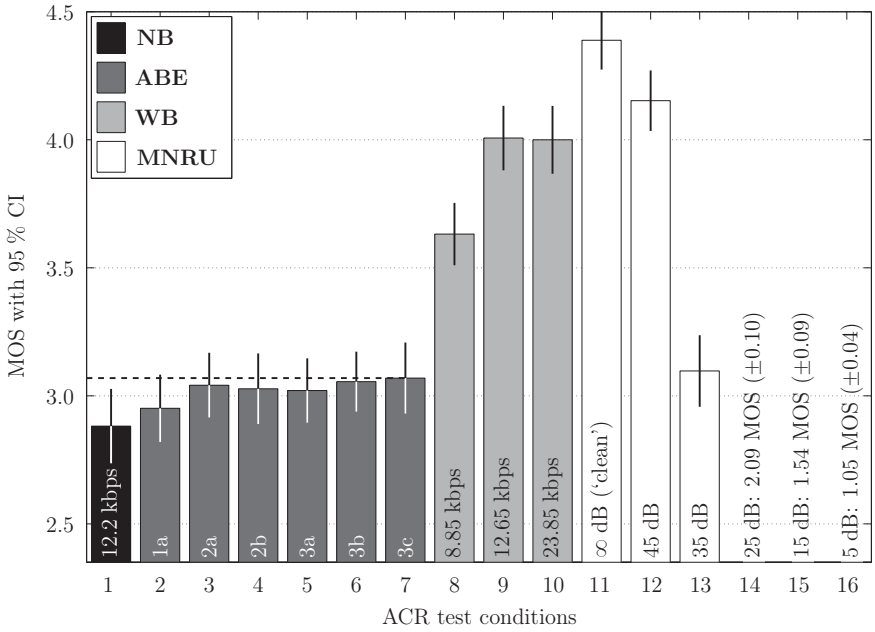


Figure 4.13: Condition-specific MOS with 95 % CI resulting from the subjective ACR listening test (the absolute numeric values are recapitulated later on in Tab. 4.7).

∞ dB, followed by the low-noise **MNRU** condition with 45 dB. Somewhat lower rated are the **WB** conditions 8–10. The lowest WB AMR bit rate turns out to be outperformed by the higher ones, which interestingly reveal a similar performance. Obviously, there is a remarkable speech quality gap of more than one MOS point between the **WB** and **NB** conditions 9 and 1, despite the comparable, widely-used WB and NB AMR bit rates of 12.65 kbps and 12.2 kbps, respectively. This offers a potential for speech quality improvement to ABE.

As expected, the MOS values attained by the **ABE** conditions 2–7 are below those of the **WB** conditions. However, they are all superior to the **NB** condition. The absolute and relative MOS gains of the **ABE** conditions compared to the **NB** condition are shown in Tab. 4.4. On the one hand, the reference ABE implementation 1a reveals a slight speech quality improvement over **NB** of 0.07 MOS points absolute and 2.41 % relative. On the other hand, the proposed ABE implementations 2a–3c are able to improve the speech quality of the **NB** condition by 0.14 . . . 0.19 MOS points absolute and 4.82 . . . 6.51 % relative. By combining the ANN-based /s/- and /z/-classification with the ANN-based UB energy adaptation, the proposed ABE implementation 3c attains the highest speech quality gains of 0.19 MOS points absolute and 6.51 % relative compared to the **NB** condition.

ABE conditions	Absolute MOS gain vs. **NB**	Relative MOS gain vs. **NB**
1a	+0.07	+2.41 %
2a	+0.16	+5.54 %
2b	+0.15	+5.06 %
3a	+0.14	+4.82 %
3b	+0.17	+6.02 %
3c	+0.19	+6.51 %

Table 4.4: Absolute and relative MOS gains of the **ABE** conditions compared to **NB**.

To statistically analyze the significance of the ACR test results concerning the **NB** and **ABE** conditions, a commonly used t-test is conducted following (ITU-T G.729EV, 2005, Annex C.3.1-2), as described in App. B.1. It shall be thereby investigated, which MOS result of the **ABE** conditions is significantly better than the one of the **NB** condition, taking into account confidence levels (CLs) of 95 % and 99 %. Tab. 4.5 shows the results of this t-test. As expected, all **ABE** conditions turn out to be not worse than **NB** for both CLs. The hypothesis of being better than the **NB** condition is fulfilled in case of a 95 % CL by all **ABE** conditions except for the reference ABE implementation 1a. Given a CL of 99 %, however, exclusively the proposed ABE implementation 3c is found to be significantly better than the **NB** condition.

CCR Test Setup

The subsequently conducted CCR listening test largely follows (ITU-T P.800, 1996, Annex E). In addition to the proposed ABE candidate 3c revealing the best ACR test performance as well as the reference ABE candidate 1a, it includes a **NB** and **WB** condition. By selecting the commonly used bit rates 12.2 kbps and 12.65 kbps for the NB AMR and WB AMR speech codec, they serve as a realistic lower and upper telephone speech quality bound, respectively. The mutual comparisons of all four telephony conditions, which result in six CCR test combinations, are rated by means of a 7-point listening-quality scale from −3 (much worse) to +3 (much better) and averaged yielding a comparison mean opinion score (CMOS) (ITU-T P.800, 1996, Annex E).

Six female as well as ten male German students with a mean age below 30 years and normal hearing abilities take part in the CCR listening test. All 16 predominantly non-expert listeners are compensated for participation by a service charge of 10 € for about 45 min instruction, familiarization, and actual test. Two listening panels with eight test sessions each are organized, where the subjects are separately assigned to. Before each test session, the listeners are briefly instructed according to (ITU-T P.800, 1996, Annex E).

t-test hypotheses	CL	**ABE** conditions					
		1a	2a	2b	3a	3b	3c
Not worse than	95 %	✓	✓	✓	✓	✓	✓
NB condition	99 %	✓	✓	✓	✓	✓	✓
Better than	95 %	✗	✓	✓	✓	✓	✓
NB condition	99 %	✗	✗	✗	✗	✗	✓

Table 4.5: Statistical t-test analysis of MOS results to check, which of the **ABE** conditions is not worse or even better than the **NB** condition for a 95 % and 99 % CL.

For familiarization purposes, the same four utterances as in the ACR listening test are spent involving all six CCR test combinations. From the resulting 24 pairwise file comparisons twelve are taken for every listener. It is thereby assured that each file pair is assigned to eight different listeners and each CCR test combination is considered twice per listener.

All 72 pairwise file comparisons deriving from the remaining twelve utterances and six CCR test combinations are dedicated to the actual test. They are equally divided into both listening panels allowing for a balanced occurrence of speaker-specific utterances and CCR test combinations. To prevent order effects, all file pairs of the actual test are randomly presented in both orders.

For each test session, the respective speech files of the familiarization phase and actual test are extracted file by file from the concatenated signals of the ACR listening test. Thus, exactly the same file-based SNR is provided in the CCR listening test. The resulting files are diotically presented in a quiet room via high-quality `AKG K 271 MK II` headphones using a laptop with a `MATLAB GUI` software and an `RME Fireface 400` external sound card. A comfortable sound level can be individually adjusted during the familiarization phase. Furthermore, the listeners are allowed to repeat single files before giving their vote. Empirically, the test sessions do not take longer than 45 min for instruction, familiarization, and actual test.

CCR Test Results

Tab. 4.6 shows the overall CMOS results of the CCR listening test, each with a 95 % CI being defined according to App. B.1. The **WB** condition expectedly outperforms the other telephony conditions. The proposed **ABE** condition 3c is rated 1.54 CMOS points worse, whereas the reference **ABE** condition 1a and the **NB** condition reveal a performance degradation of 1.75 and 2.16 CMOS points, respectively. In direct comparison, **ABE** condition 3c performs 0.34 CMOS points better than **ABE** condition 1a. Obviously, both **ABE** conditions are superior to the **NB** condition. On the one hand, **ABE** 1a achieves an improvement

CCR test combinations		CMOS with 95 % CI
NB @12.2 kbps	**WB** @12.65 kbps	−2.16 (±0.11)
ABE 1a	**WB** @12.65 kbps	−1.75 (±0.11)
ABE 3c	**WB** @12.65 kbps	−1.54 (±0.13)
ABE 3c	**ABE** 1a	+0.34 (±0.10)
ABE 1a	**NB** @12.2 kbps	+0.88 (±0.11)
ABE 3c	**NB** @12.2 kbps	+1.01 (±0.13)

Table 4.6: CMOS results of the subjective CCR listening test: A positive CMOS indicates that the first column is rated better than the second one, and vice versa.

of 0.88 CMOS points. On the other hand, **ABE** 3c is capable of improving the speech quality compared to the **NB** condition by 1.01 CMOS points, which is almost half as much as the gain of 2.16 CMOS points attained by the **WB** condition. The proposed ABE candidate 3c is therefore selected as the final recommendation of this work.

4.3.2 Instrumental Speech Quality Measurements

In contrast to the subjective speech quality perception serving as the ground truth for the given human-to-human ABE application, a speech quality prediction via instrumental measurements saves time and costs. Intrusive (i.e., reference-based) assessment methods requiring both the modified and reference signals generally attain higher accuracies than non-intrusive (i.e., reference-free) ones (Pulakka, 2013, Sec. 6.2.2). They commonly rely on simple distance measures or more sophisticated models predicting the speech quality as perceived by humans (Pulakka, 2013, Sec. 6.2).

Among the numerous distance measures (Iser et al., 2008, Sec. 3.4), the LSD has been widely used for evaluation of ABE systems (Pulakka, 2013, Sec. 6.2.1). In this work, a WB LSD based on (3.6) is employed by adapting k_{low} and k_{high} to the frequency range of $0 \ldots 8$ kHz. Thus, it aims at accurately predicting the subjectively perceived speech quality based on the full bandwidth.

Instrumental assessment methods that explicitly model the human speech quality perception are basically divided into two types. On the one hand, the overall perceptual speech quality can be predicted by combining several quality dimensions (Pulakka, 2013, Sec. 6.2.3). They may include the coloration, noisiness, discontinuity, and loudness, such as in case of the diagnostic instrumental assessment of listening quality (DIAL) measure being developed from the telecommunication objective speech quality assessment (TOSQA) core model (Côté, 2011). It has been utilized in the context of ABE evaluation in (Möller et al., 2013). On

the other hand, a direct prediction of the overall perceptual speech quality can be attained via the well-known perceptual evaluation of speech quality (PESQ) (ITU-T P.862, 2001) adopting an MOS-listening quality objective (LQO) mapping (ITU-T P.862.1, 2003; ITU-T P.800.1, 2006) as well as a WB extension (ITU-T P.862.2, 2007). WB PESQ has been widely applied for ABE evaluation purposes (Pulakka, 2013, Sec. 6.2.2). Furthermore, its successor perceptual objective listening quality assessment (POLQA) can be employed (ITU-T P.863, 2011). Compared to WB PESQ, it is also capable of evaluating super-WB and fullband speech signals with a sampling rate until 48 kHz and includes some further innovations. WB PESQ and POLQA have been jointly investigated for the purpose of ABE evaluation in (Möller et al., 2013; Bauer et al., 2014c). They were found to capture ABE somewhat better than DIAL by (Möller et al., 2013) and are therefore employed in this work. Please note that WB PESQ is applied by activating the WB mode and setting the sampling rate to 16 kHz (ITU-T P.862.2, 2007). For POLQA, the super-WB mode is activated and no fixed active speech level used (ITU-T P.863, 2011; ITU-T P.863.1, 2013).

WB PESQ, POLQA and WB LSD Measurement Results

Tab. 4.7 recapitulates the subjective ACR test results. Furthermore, it shows the respective instrumental predictions resulting from the WB PESQ, POLQA and WB LSD measurements averaged over the single telephony conditions. For all intrusive assessments, the **MNRU** condition at ∞ dB provides the corresponding reference signals. Please note that the MOS unit of the subjective ACR test results has been complemented in Tab. 4.7 by the specification listening quality subjective (LQS) to better distinguish them from the MOS-LQO measurements according to (ITU-T P.800.1, 2006). Apart from that, the WB LSD results characterize a distortion measured in dB.

Both WB PESQ and POLQA attain meaningful rank orders of the **MNRU** and **WB** conditions. By increasing the ratio of speech to modulated noise power or the WB AMR bit rate, higher MOS-LQO values are obtained. Please note that the highest possible MOS-LQO of both WB PESQ and POLQA is expectedly obtained by the **MNRU** condition at ∞ dB serving as reference. In contrast, the **MNRU** condition at 5 dB performs worst. Despite the correct rankings, there are sometimes relatively high absolute deviations from the MOS-LQS results. In case of the **MNRU** condition at 35 dB, e.g., WB PESQ and POLQA drastically overestimate the ground truth by 0.68 and 1.26 MOS points, respectively. Moreover, they underestimate the gap between the **WB** condition at 12.65 kbps and the **NB** condition at 12.2 kbps considerably and therefore dedicate less potential for speech quality improvement to ABE. When focusing only on the **NB** and **ABE** conditions, neither WB PESQ nor POLQA meets the rank order of the subjective test results correctly. On the one hand, WB PESQ rates the **NB** condition slightly better than the **ABE** conditions 2a and 3a.

Telephony conditions		ACR test [MOS-LQS]	WB PESQ [MOS-LQO]	POLQA [MOS-LQO]	WB LSD [dB]
NB	12.2 kbps	2.88 (±0.15)	3.03	3.51	15.70
ABE	1a	2.95 (±0.13)	3.15	3.66	12.12
	2a	3.04 (±0.13)	2.99	3.65	11.96
	2b	3.03 (±0.14)	3.18	3.70	12.63
	3a	3.02 (±0.13)	2.99	3.67	10.84
	3b	3.06 (±0.12)	3.11	3.68	11.66
	3c	3.07 (±0.14)	3.08	3.70	11.60
WB	8.85 kbps	3.63 (±0.12)	3.29	3.79	8.47
	12.65 kbps	4.01 (±0.13)	3.64	4.15	7.94
	23.85 kbps	4.00 (±0.13)	3.83	4.45	7.35
MNRU	∞ dB	4.39 (±0.11)	4.64	4.70	0.00
	45 dB	4.15 (±0.12)	4.08	4.66	5.90
	35 dB	3.10 (±0.14)	3.78	4.36	6.83
	25 dB	2.09 (±0.10)	2.75	2.76	8.63
	15 dB	1.54 (±0.09)	1.57	1.44	11.37
	5 dB	1.05 (±0.04)	1.15	1.18	15.28

Table 4.7: Intrumental predictions of the subjective ACR test results (the 95 % CIs specified in brackets are not being predicted) using WB PESQ, POLQA and WB LSD measurements.

Additionally, it judges **ABE** 2b as the best ABE candidate, followed by **ABE** 1a, **ABE** 3b, and the actual winner **ABE** 3c. On the other hand, the **NB** condition is indeed consistently ranked worse than all **ABE** conditions by POLQA, however, the ranking among the ABE candidates is wrong. At least, the **ABE** condition 3c reveals the best performance along with **ABE** 2b, being followed by **ABE** 3b, **ABE** 3a, **ABE** 1a, and **ABE** 2a. Obviously, the **ABE** conditions are predicted by POLQA more than half MOS-LQO point higher than by WB PESQ, with WB PESQ approximating the absolute MOS-LQS values much better. Furthermore, the POLQA predictions of all ABE candidates lie within a relatively small range between 3.65 and 3.70 MOS-LQO.

The WB LSD measurement results in dB are not directly comparable with the absolute MOS-LQS and MOS-LQO values. Nevertheless, they reversely reflect the same rank orders of the **MNRU** and **WB** conditions as WB PESQ and POLQA. As expected, the **MNRU** condition at ∞ dB reveals a WB LSD of 0.00 dB, since the modified and reference signals for the intrusive distance computation are equal in this case. Interestingly, the **NB** condition is outperformed by all other telephony conditions. Hence, the WB LSD measure punishes the lack of UB frequencies stronger than WB PESQ and POLQA, which capture the human

Correlation analysis		WB PESQ	POLQA	WB LSD
Over all telephony conditions	RMSE [MOS points]	0.29	0.59	-
	Pearson correlation	+0.95	+0.95	−0.68
	Spearman rank corr.	+0.90	+0.96	−0.76
Over all **MNRU** conditions	RMSE [MOS points]	0.40	0.63	-
	Pearson correlation	+0.97	+0.96	−0.93
	Spearman rank corr.	+1.00	+1.00	−1.00
Over all **NB** and **WB** conditions	RMSE [MOS points]	0.28	0.40	-
	Pearson correlation	+0.92	+0.90	−0.97
	Spearman rank corr.	+0.80	+0.80	−0.80
Over all **ABE** conditions	RMSE [MOS points]	0.11	0.65	-
	Pearson correlation	−0.32	+0.51	−0.22
	Spearman rank corr.	−0.26	+0.46	−0.31
Over all **ABE** conditions as well as **NB** @12.2 kbps and **WB** @12.65 kbps	RMSE [MOS points]	0.17	0.61	-
	Pearson correlation	+0.93	+0.98	−0.82
	Spearman rank corr.	+0.33	+0.78	−0.71

Table 4.8: RMSE, Pearson correlation, and Spearman rank correlation between instrumental measurements and subjective ACR test results for different sets of telephony conditions.

speech quality perception in this respect better. This offers to ABE a relatively big potential for speech quality improvement. Although, it is exploited best by the finally optimized ABE implementations 3a–3c, the rank order does not match with the subjective test results and the rankings of the other **ABE** conditions neither agree.

Correlation Analysis of Instrumental and Subjective Results

By means of the correlation analysis described in App. B.2, the instrumental measurements are systematically compared with the subjective ACR test results taking into account a root mean square error (RMSE), Pearson correlation, and Spearman rank correlation. Tab. 4.8 shows the corresponding results for different sets of telephony conditions.

Due to the mismatch between dB and MOS, the RMSE is not evaluated in case of the WB LSD measure. WB PESQ obviously attains a lower RMSE than POLQA for all sets of telephony conditions. The corresponding RMSEs based on all telephony conditions result in 0.29 and 0.59 MOS points, respectively. When focusing on the **ABE** conditions, WB PESQ outperforms POLQA by 0.54 MOS points. This absolute RMSE difference slightly shrinks to 0.44 MOS points, if the **NB** condition at 12.2 kbps and the **WB** condition at 12.65 kbps are also included. Hence, WB PESQ outperforms POLQA in terms of RMSE. Moreover, it

is a disadvantage of the WB LSD measure that the MOS results are not directly predictable.

The Pearson correlation over all **ABE** conditions is very low both for WB PESQ and POLQA. While POLQA at least provides a value of +0.51, WB PESQ even reveals a negative correlation of −0.32. However, the correlation results significantly improve to +0.98 and +0.93, respectively, when adding the **NB** condition at 12.2 kbps and the **WB** condition at 12.65 kbps. The correlations for all remaining sets of telephony conditions neither fall below +0.90. Hence, it is assumed that the Pearson correlation is not fully capable of evaluating the reliability of the instrumental measurements exclusively for the **ABE** conditions. This assumption is also confirmed by the correlation results of the WB LSD measure. Please note that their negative signs correctly reflect the reverse meaning of a distortion in dB and a quality prediction in MOS. Again, the **ABE** conditions yield the worst correlation of −0.22, which is enormously improved to −0.82 through the inclusion of the **NB** condition at 12.2 kbps and the **WB** condition at 12.65 kbps. For the **NB** plus **WB** conditions as well as for the **MNRU** conditions, the correlation is lower than −0.90. When taking into account all telephony conditions, however, it is degraded to −0.68 unlike in case of WB PESQ and POLQA. As expected, the WB LSD measure turns out to be less reliable in predicting the subjective ACR test results.

A perfect Spearman rank correlation is attained by all instrumental assessment methods for the **MNRU** conditions. In fact, this would also hold for the **NB** plus **WB** conditions, if the misleading ACR test rankings of the **WB** conditions at 12.65 kbps and 23.85 kbps according to Tab. 4.7 were corrected. Based on all telephony conditions, the WB LSD measure reveals a worse rank correlation than WB PESQ and POLQA. The rank order of the **ABE** conditions is poorly predicted by all instrumental assessment methods. WB PESQ even yields a negative rank correlation of −0.26. In contrast, POLQA performs better with +0.46 followed by WB LSD with −0.31. Furthermore, only POLQA and WB LSD rank the **NB** condition at 12.2 kbps and the **WB** condition at 12.65 kbps correctly in relation to the **ABE** conditions. Taken as a whole, POLQA turns out to provide superior rank correlations than both WB PESQ and WB LSD.

4.3.3 Discussion

As mentioned at the beginning of this section, the presented speech quality assessment aims at investigating the ability of ABE for speech quality enhancement and the reliability of instrumental measurements compared to subjective listening tests for ABE evaluation.

In Sec. 4.3.1, a two-stage subjective assessment methodology is specifically developed to find out, whether ABE is capable of enhancing speech quality. Inspired by typical speech codec standardization efforts, it consists of a qualification and selection phase. These two

phases are successively realized by an ACR and CCR listening test. A significant improvement of 0.19 MOS points absolute and 6.51 % relative compared to the NB telephone speech quality is attained in the ACR test by the proposed ABE implementation 3c, which includes all ABE optimizations of Sec. 4.2. According to a statistical t-test analysis, it is found to be significantly better given a CL of 99 %. Comparable speech quality gains obtained from ACR listening tests have been rarely reported in literature so far (Ramabadran and Jasiuk, 2008; Bauer et al., 2014c; Pulakka et al., 2014). This may be explained by a relatively low ACR test sensitivity[14] with respect to ABE. Due to the successive presentation of single files, the listeners seem to individually rate each speech signal based on a balanced weighting of all perceived positive and negative impressions. This specifically means for ABE that introduced artifacts counteract the extended bandwidth (Bauer et al., 2014c). In general, a CCR test is expected to provide more sensitive results (Möller, 2000). Despite potential ABE artifacts, listeners are assumed to be biased by a direct file comparison towards the higher bandwidth (Bauer et al., 2014c). This assumption is confirmed by the CCR test results given in Tab. 4.6. They certify the reference ABE implementation 1a a remarkable gain of 0.88 CMOS points over the NB telephone speech quality, although the ACR test reveals only a negligible benefit of 0.07 MOS points following Tab. 4.4. The proposed ABE implementation 3c even points out a superior CCR test performance. It outperforms the reference ABE implementation 1a by 0.34 CMOS points and improves the NB telephone speech quality by 1.01 CMOS points. The latter improvement bridges almost half of the gap between the NB and WB telephone speech quality amounting to 2.16 CMOS points. Thus, the proposed ABE implementation 3c demonstrates a significant speech quality enhancement in the ACR and CCR test. Both are considered to be relevant for telephony scenarios in practice: While successive calls in a single NB, ABE, or WB mode are better captured by the ACR test, the CCR test simulates handover calls switching between these modes (Bauer et al., 2014c).

Sec. 4.3.2 investigates, whether subjective listening tests for ABE evaluation can be reliably replaced by instrumental measurements to save time and costs. In addition to the simple distance measure WB LSD, the more sophisticated models WB PESQ and POLQA predicting the overall perceptual speech quality are involved. They are systematically compared with the ACR test results serving as ground truth by means of an extensive correlation analysis. It takes into account the RMSE, Pearson correlation, and Spearman rank correlation. Since the WB LSD is measured in dB and does therefore not directly predict MOS results, it is not evaluated by an RMSE. Apart from that, WB PESQ outperforms POLQA in terms of RMSE. Focusing on the Pearson correlation, the WB LSD measure turns out to be less reliable than WB PESQ and POLQA. However, the correlation results point out inconsistencies, when restricting the evaluation exclusively to ABE results. This special case

[14]Please note that Pulakka et al. (2014) successfully used a 9-point instead of the 5-point listening-quality scale for rating to obtain more sensitive ACR test results.

is captured better by the Spearman rank correlation. All instrumental assessment methods have problems in predicting the correct rank order of the ABE candidates. Nevertheless, POLQA provides a superior rank correlation than WB PESQ and WB LSD. Taking into account the complete correlation analysis, WB PESQ and POLQA expectedly turn out to be more reliable than the WB LSD measure. However, neither WB PESQ nor POLQA can fully replace the use of subjective listening tests. This conclusion confirms the findings in (Möller et al., 2013; Fingscheidt and Bauer, 2013; Bauer et al., 2014c).

4.4 Summary

In this chapter, the ABE framework exploiting phonetic *a priori* knowledge is further developed to enhance NB telephone speech services. Due to the online requirements of this human-to-human ABE application, phoneme class labels are only available for offline ABE training.

Even without phonetic support of the ABE processing, the preliminary syllable articulation tests in Sec. 4.1 reveal that ABE is capable of significantly improving the NB telephone speech intelligibility in all cases involving a low SNR, language mismatch, and hearing impairment, according to Tab. 4.1. Among the tested critical fricatives, the phonemes /s/ and /z/ particularly benefit from the use of ABE.

Before investigating the impact of ABE on the NB telephone speech quality, under- and overestimation artifacts need to be suppressed. For this purpose, Sec. 4.2 introduces the ABE optimizations highlighted in Fig. 4.5. By means of a phonetic ANN classifier, phoneme class labels of the crucial /s/- and /z/-sounds are obtained in real time to support the HMM-based estimation process. A reduction of underestimation artifacts is thereby achieved, however, at the expense of overestimating some other phonemes. Hence, a second ANN classifier is employed to adaptively correct the overestimated UB energy in the cepstral domain. In case of a failure during speech pauses, which can be easily confused with /s/- and /z/-sounds based on the LB spectrum, a robust SPD serves to suppress overestimations in the residual domain. Apart from that, transient and discontinuity artifacts may arise from the time-variant LP analysis and synthesis filtering. They are hardly audible but clearly visible in the spectrograms of Fig. 4.10. These switching effects are reduced by temporally smoothing the estimated WB LP filter coefficients after frame transitions. Furthermore, the alias distortions visualized in the spectrograms of Fig. 4.11 are removed by means of an anti-aliasing highpass filter. For this purpose, a conversion of the serial into a parallel ABE structure is required, as depicted in Fig. 4.5.

The final speech quality assessment in Sec. 4.3 points out that the optimized ABE algo-

rithm is capable of significantly enhancing the NB telephone speech quality. This is proven by ACR and CCR listening tests in the context of a specifically developed two-stage subjective assessment methodology, following Tab. 4.4–4.6. Instrumental measurements based on WB PESQ, POLQA, and WB LSD are also conducted to investigate, whether they can replace the time-consuming and cost-intensive subjective listening tests. However, a systematic correlation analysis including the RMSE, Pearson correlation, and Spearman rank correlation reveals in Tab. 4.8 that none of the instrumental assessment methods is able to reliably predict the ACR test results serving as ground truth.

Chapter 5

Conclusions and Outlook

In this work, a statistical framework for high-band ABE based on the state-of-the-art HMM approach of Jax (2002) has been formulated. Moreover, it has been developed further by several algorithmic innovations with the objective of improving speech intelligibility and quality. An exploitation of phonetic *a priori* knowledge in support of both ABE training and processing thereby represents the most important innovation. It aims at reducing artifacts that typically arise from ABE due to the confusion of critical phonemes. The phonetic *a priori* knowledge is provided in terms of frame-wise phoneme class labels. They allow for a *supervised* CB training and thereby implicitly support also the subsequently trained LDA as well as HMM. Within ABE processing, the frame-wise phoneme class labels are integrated into the HMM decoder by means of a novel phoneme class probability matrix modifying the observation likelihoods. Further innovations comprise an integration of the following techniques into the ABE framework: A non-rectangular windowing with window overlap, a VA-based optimal state sequence decoder, a cepstral smoothing strategy, as well as an additional control over the UB energy and cut-off frequency.

After the corresponding technological fundamentals have been established, the *two main ABE use cases in practice* are addressed. On the one hand, the use of ABE for the training of WB telephony ASR systems reflects a human-to-machine ABE application without online requirements. On the other hand, the use of ABE for the enhancement of NB telephone speech services represents a human-to-human ABE application with online requirements.

Offline ABE for Training of WB Telephony ASR Systems

Recognition performance benefits from an utilization of more speech data for ASR training, particularly when dealing with challenging recognition tasks (Church and Mercer, 1993). IVR systems supporting HD telephony services therefore need to be trained on huge amounts of WB telephone speech material. However, there is a lack of WB telephone speech corpora.

This problem is tackled by making use of existing NB telephone speech databases. In order to prepare them for the training of WB telephony ASR systems, they can be upgraded in speech bandwidth via high-band ABE. Expensive and time-consuming speech recordings are thereby prevented. Furthermore, such an ABE-based speech database upgrade works completely independent from the employed automatic speech recognizer, i.e., no changes of the ASR system are required. As this human-to-machine ABE application does not demand any online requirements, the phonetic *a priori* information is or can be made available offline for both ABE training and processing.

Interestingly, the phoneme recognition performance turns out to hardly depend on the speech bandwidth for most of the phonemes. In case of a larger speech bandwidth, exclusively the fricatives /s/ and /z/ reveal a considerable gain in recognition rate. Thus, they need to be taken more into account by ABE than the remaining phonemes. Based on these findings, several phonetically motivated CBs for ABE have been designed using diversified phoneme classes. The improved spectral reconstruction with specifically trained CB representatives mainly aims at reducing lisping artifacts that typically arise from ABE due to an underestimation of /s/- and /z/-sounds. However, this comes along with HMM over-representations of /s/- and /z/-states. They particularly affect other fricatives, such as /f/ and /th/, as well as speech pauses, which – based on the LB spectrum – can be easily confused with /s/ and /z/. These phonetic confusions provoke temporal smearing effects. To suppress them, a three-stage modification of the state transition probability matrix has been developed. It involves an appropriate smoothing, attenuation and boosting of the HMM state transitions. From these modifications, the fricatives /s/, /z/, /f/, and /th/ as well as the speech pauses significantly benefit. Moreover, none of the remaining phonemes suffers.

To investigate the usability of the ABE-based speech database upgrade for WB telephony ASR training, practice-relevant large-vocabulary recognition experiments have been conducted. Before applying the ABE, relevant ASR baseline results have been simulated for the purpose of comparison. In the given showcase, the WER increases by 6 % absolute, when quartering the amount of WB telephone speech training data. By filling up the missing three quarters of training data with interpolated NB speech, the absolute WER increase of 6 % decreases to 3.12 %. This ASR baseline result reflects a lower ABE performance bound, assuming interpolation to be the most rudimentary form of ABE. It is reduced by 1.9 % absolute via an ideal ABE taking the excitation and spectral envelope from the original WB speech signal which is, however, not available in practice. Anyway, the further developed offline ABE proposed in this work attains a respectable WER reduction of 1.12 % (instead of 1.9 %) absolute. All in all, it turns out to minimize the absolute WER distance between the WB ASR baselines with full amount and one quarter of training data from 6 % to only 2 %. This leads to the conclusion that an offline ABE upgrade of NB telephone speech databases

is recommendable in case of insufficient training data for WB telephony ASR systems.

Online ABE for Enhancement of NB Telephone Speech Services

Network- as well as provider-specific restrictions and incompatibilities are still the reason why phone calls are oftentimes established in NB mode. Due to its limited acoustic bandwidth, NB telephony reveals a reduced speech intelligibility and quality. High-band ABE aims at enhancing both by estimating and reconstructing the missing upper frequency components. But it tends to confuse particularly the fricatives /s/ and /z/ with other phonemes in case of ambiguous spectral characteristics based on the NB frequency range. Annoying under- and overestimation artifacts are thereby provoked. As mentioned before, these problems can be tackled by means of phonetic *a priori* knowledge. However, this information is or can be made available offline only for ABE training, since the underlying human-to-human application requires an online ABE processing without any access to phonetic *a priori* knowledge.

It has already been shown that the further developed ABE can be successfully employed offline for ASR training purposes with the objective of increasing speech recognizer performance. To investigate, whether human speech *intelligibility* can be improved online by ABE, extensive syllable articulation tests have been performed. Based on the test results it can be concluded that hearing-impaired as well as non-native listeners benefit from online ABE particularly during /s/- and /z/-sounds. Native, normal-hearing listeners turn out to profit, too, but only in noisy environments. Nevertheless, severe artifacts were partially noticeable in the syllable articulation tests. It became thereby clear that the online ABE needs to be further optimized, when trying to enhance telephone speech *quality* from a human point of view.

Several optimizations have been developed addressing the under- and overestimation artifacts. An ANN-based real-time classification of the crucial /s/- and /z/-sounds represents the most challenging innovation. It provides the missing phonetic *a priori* knowledge required for online ABE processing. Underestimation artifacts are thereby successfully reduced at the expense of additional overestimation artifacts. A second ANN classifier is therefore employed to adaptively correct the overestimated UB energy in the cepstral domain. Due to the fact that speech pauses can be easily confused with /s/- and /z/-sounds based on the LB spectrum, a robust SPD is applied to suppress overestimations in the residual domain. During frame transitions, transients as well as discontinuities potentially arising from time-variant LP analysis and synthesis filtering are removed by temporally smoothing the estimated WB LP filter coefficients. Finally, a serial-to-parallel conversion of the ABE structure has been done to eliminate alias distortions in the LB spectrum of the estimated UB speech signal via anti-aliasing highpass filtering.

The ability of the optimized online ABE for enhancing telephone speech quality from a human perspective has been extensively assessed by widely recognized subjective listening tests and instrumental measurements. Inspired by typical speech codec standardization efforts, a two-stage subjective assessment methodology has been developed. It involves an initial qualification phase realized by an ACR listening test and a subsequent selection phase represented by a CCR listening test. The ACR test results point out that the proposed ABE improves the NB telephone speech quality by 0.19 MOS points absolute and 6.51 % relative, respectively. By means of a statistical t-test analysis, the significance of this improvement has been successfully confirmed given a 99 % CL. The CCR test results reveal a remarkable quality gain of 1.01 CMOS points, when enhancing the NB telephone speech by the proposed ABE. It bridges almost half of the gap between the NB and WB telephone speech quality. The instrumental measurements based on WB PESQ, POLQA, and WB LSD have not been able to reliably predict the subjective results of the ACR test. This has been confirmed by means of a systematic correlation analysis including the RMSE, Pearson correlation, and Spearman rank correlation. Hence, time-consuming and cost-intensive subjective listening tests can still not be replaced by instrumental measurements at least for ABE assessment purposes.

The overall test results lead to the conclusion that the online ABE proposed in this work is capable of significantly enhancing NB telephone speech services in terms of human speech intelligibility and quality. Due to its moderate algorithmic delay of about 25 ms (including the frame length $N = 10$ ms) and computational complexity around 35 MFLOPS (i.e., for the worst observed frame), it can be employed for online applications in practice (Bauer et al., 2010b).

Future Challenges of ABE

Beyond the scope of this work, several challenges are still to be met by ABE research in the future. Amongst others, the latest advancements in the field of deep and recurrent ANNs may yield further improvements on high-band ABE performance. However, this requires the possibility of time- and cost-efficient ABE evaluation. A reliable instrumental ABE measure therefore needs to be developed to get finally rid of extensive subjective listening tests. The phonetic dependencies of ABE should be thereby taken into account, as proposed in (Fingscheidt and Bauer, 2013). Apart from that, low-band ABE still bears a great potential for further speech quality enhancements. Until now, it could not be fully exploited due to the partially inaccurate pitch estimation and reconstruction during noisy speech parts provoking severe artifacts. Although *HD Voice* services reach a more and more complete market penetration, at least in mobile telephony, ABE will still play a role, e.g., for the purpose of WB to super-WB extension.

Appendix A

Phoneme Alphabet SAMPA-D-VMlex

Phoneme group	Phoneme	Example	
		Word	Phonetic transcription
Fricatives	/s/	da**s**	/d a s/
	/z/	**s**ein	/z aI n/
	/S/	**Sch**ein	/S aI n/
	/Z/	Eta**g**e	/E t a: Z @/
	/f/	**f**ast	/f a s t/
	/v/	**w**as	/v a s/
	/w/	Jan**u**ar	/j a n w a: r/
	/C/	ri**ch**tig	/r I C t I k/
	/x/	no**ch**	/n O x/
	/h/	**H**err	/h E 6/
	/r/	**R**eise	/r aI z @/
Plosives	/p/	Urlau**b**	/u: r l aU p/
	/b/	**b**is	/b I s/
	/t/	is**t**	/? I s t/
	/d/	**d**er	/d E 6/
	/k/	Zu**g**	/t s u: k/
	/g/	**g**leich	/g l aI C/
Sonorant consonants	/j/	**j**etzt	/j E t s t/
	/m/	**m**ir	/m i: r/
	/n/	e**n**dlich	/E n t l I C/
	/N/	Da**n**k	/d a N k/
	/l/	verp**l**ant	/f E r p l a: n t/

Vowels	/I/	**i**nnerhalb	/I n 6 h a l p/
	/i:/	z**ie**hen	/t s i: @ n/
	/Y/	m**ü**ssen	/m Y s @ n/
	/y:/	**ü**ber	/? y: b 6/
	/9/	k**ö**nnte	/k 9 n t @/
	/2:/	k**ö**nnen	/k 2: n @ n/
	/9∼/	Parf**ü**m	/p a f 9∼ m/
	/E/	schl**e**cht	/S l E C t/
	/E:/	betr**ä**gt	/b @ t r E: k t/
	/e:/	**e**wig	/? e: v I C/
	/e∼/	T**ein**t	/t e∼ n/
	/U/	**u**nd	/? U n t/
	/u:/	**U**hr	/? u: 6/
	/O/	**O**rdnung	/? O 6 t n U N/
	/o:/	M**o**ntag	/m o: n t a: k/
	/o∼/	Sais**on**	/s E: s o∼ n/
	/a/	w**a**s	/v a s/
	/a:/	Donnerst**a**g	/d O n 6 s t a: k/
	/a∼/	Restaur**ant**	/r E s t o: r a /
	/6/	fü**r**	/f y: 6/
	/@/	fahr**e**n	/f a: r @ n/
Diphthongs	/aI/	b**ei**	/b aI/
	/aU/	**au**ch	/? aU x/
	/OI/	n**eu**n	/n OI n
Glottal stop	/?/	The-ater	/t E ? a: t 6/
Silent/distorted speech pauses	/si/	-	-
	/VN/	-	-
	/NS/	-	-

Table A.1: Phoneme alphabet SAMPA-D-VMlex with phoneme group assignments and exemplary phonetic transcriptions according to (J. Abel, 2013, App. A).

Appendix B

Statistical Analysis

The following confidence and correlation analysis plays an important role to statistically analyze the significance of results obtained in this work. For generalization purposes, two random samples $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$ of size N are defined. They are assumed to arise from normal distributions with unknown expectations μ_x, μ_y and unknown variances σ_x^2, σ_y^2. The respective arithmetic means $\hat{\mu}_x$, $\hat{\mu}_y$ and standard deviations $\hat{\sigma}_x$, $\hat{\sigma}_y$ can be derived from both random samples as follows (Ross, 2006, Sec. 2.3):

$$\hat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \hat{\sigma}_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}_x)^2}, \tag{B.1}$$

$$\hat{\mu}_y = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad \hat{\sigma}_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{\mu}_y)^2}. \tag{B.2}$$

B.1 Confidence Analysis

According to (Ross, 2006, Sec. 7.3.1), the unknown expectations of the underlying normal distributions are assumed to lie in between the following confidence intervals depending on a confidence level of $CL \in (0,1)$ and $N-1$ degrees of freedom:

$$\hat{\mu}_x - \frac{\hat{\sigma}_x}{\sqrt{N}} \cdot t_{CL,N-1} < \mu_x < \hat{\mu}_x + \frac{\hat{\sigma}_x}{\sqrt{N}} \cdot t_{CL,N-1}, \tag{B.3}$$

$$\hat{\mu}_y - \frac{\hat{\sigma}_y}{\sqrt{N}} \cdot t_{CL,N-1} < \mu_y < \hat{\mu}_y + \frac{\hat{\sigma}_y}{\sqrt{N}} \cdot t_{CL,N-1}. \tag{B.4}$$

The corresponding t-value $t_{CL,N-1}$ needs to be taken from a look-up table of the *two-sided* Student's t-distribution, e.g., in (Federighi, 1959). For the sake of convenience, the 95 % CIs in Fig. 4.2, Fig. 4.13, and Tab. 4.6 have been computed by setting $t_{0.95,N-1} = 1.96$. This approximation is valid due to the numerous degrees of freedom involved (i.e., $N-1 \geqslant 95$).

To investigate, whether an **ABE** condition performs significantly better than the **NB** condition regarding the subjective ACR listening test in Sec. 4.3.1, the widely spread t-tests have been conducted following (ITU-T G.729EV, 2005, Annex C.3.1-2). For this purpose, random sample $x_1, x_2, \ldots, x_N$ shall represent the speaker- as well as listener-specific MOS results of the **NB** condition, and $y_1, y_2, \ldots, y_N$ those of a particular **ABE** condition. As the *common* standard deviations $\hat{\sigma}_x$ and $\hat{\sigma}_y$ only depend on their respective random sample, the simple detection of an overlap between the CIs in Eqs. (B.3)–(B.4) does not necessarily disprove a potentially significant difference between μ_x and μ_y (e.g., between the **NB** condition 1 and the **ABE** condition 7 in Fig. 4.13). A *pooled* standard deviation is therefore specified taking both random samples into account (ITU-T G.729EV, 2005, Annex C.3.1)

$$\hat{\sigma}_{x-y} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left((x_i - y_i) - \hat{\mu}_{x-y}\right)^2}, \tag{B.5}$$

with the arithmetic mean of the pairwise MOS differences given by

$$\hat{\mu}_{x-y} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i). \tag{B.6}$$

The applied paired t-tests are assumed to benefit from this random sample pooling. On the one hand, the null hypothesis $\mu_y \geqslant \mu_x$ (i.e., the expected MOS of the **ABE** condition is *not worse than* the one of the **NB** condition) can be verified by means of a *one-sided* t-test. A confirmation of the null hypothesis is attained, if the following inequation holds (ITU-T G.729EV, 2005, Annex C.3.1)

$$\hat{\mu}_y \geqslant \hat{\mu}_x - \frac{\hat{\sigma}_{x-y}}{\sqrt{N}} \cdot t_{CL,N-1}. \tag{B.7}$$

On the other hand, the null hypothesis $\mu_y > \mu_x$ (i.e., the expected MOS of the **ABE** condition is *better than* the one of the **NB** condition) can be verified by another *one-sided* t-test. This time, the null hypothesis is confirmed in case of the inequation (ITU-T G.729EV, 2005, Annex C.3.2)

$$\hat{\mu}_y > \hat{\mu}_x + \frac{\hat{\sigma}_{x-y}}{\sqrt{N}} \cdot t_{CL,N-1}. \tag{B.8}$$

Due to the combination of 4 speakers and 24 listeners, the t-test results based on CLs of $CL \in \{0.95, 0.99\}$ in Tab. 4.5 consider $N-1 = 4 \cdot 24 - 1 = 95$ degrees of freedom. Following (Federighi, 1959), the t-values have been therefore set to $t_{0.95,95} = 1.661$ and $t_{0.99,95} = 2.366$, respectively.

B.2 Correlation Analysis

To evaluate, how reliable the instrumental measurements predict the subjectively perceived speech quality, a correlation analysis has been conducted in Sec. 4.3.2. For this purpose,

random sample $x_1, x_2, \ldots, x_N$ shall represent the file-based instrumental measurement results (i.e., MOS-LQO for PESQ and POLQA, respectively, or WB LSD in dB). Furthermore, MOS-LQS results of the subjective ACR listening test are obtained by averaging all listener-specific ratings per file to serve as random sample $y_1, y_2, \ldots, y_N$.

According to (ITU-T P.1401, 2012), the MOS-based results have been compared (i.e., MOS-LQO vs. MOS-LQS) by means of the commonly used RMSE measure (Côté, 2011)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}. \tag{B.9}$$

Unfortunately, it is not applicable to the WB LSD results because of the mismatching units dB and MOS. Furthermore, the RMSE implies a relatively simple error criterion. To avoid these restrictions, two more sophisticated correlation measures have been employed. Compared to the RMSE, the well-known Pearson correlation yields a dimensionless coefficient normalized to the interval $[-1, 1]$ (ITU-T P.862, 2001, Sec. 7.1)

$$CORR_{\text{Pearson}} = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \hat{\mu}_x)^2 \sum_{i=1}^{N}(y_i - \hat{\mu}_y)^2}}. \tag{B.10}$$

A Pearson correlation of 0 indicates that the involved random samples do not linearly depend on each other at all. Minor modifications of Eq. (B.10) directly lead to the Spearman rank correlation that focuses on the ranking of the random samples. Its coefficient is defined as (Iser et al., 2008, Sec. 6.1.3)

$$CORR_{\text{Spearman}} = \frac{\sum_{i=1}^{N}\left(\text{rank}(x_i) - \overline{\text{rank}}_x\right)\left(\text{rank}(y_i) - \overline{\text{rank}}_y\right)}{\sqrt{\sum_{i=1}^{N}\left(\text{rank}(x_i) - \overline{\text{rank}}_x\right)^2 \sum_{i=1}^{N}\left(\text{rank}(y_i) - \overline{\text{rank}}_y\right)^2}}, \tag{B.11}$$

with $\text{rank}(x_i)$, $\text{rank}(y_i)$ denoting the respective rank of x_i, y_i. Averaging over all ranks of the respective random sample yields $\overline{\text{rank}}_x$, $\overline{\text{rank}}_y$.

Please note that all three measures can be flexibly applied to certain groups of telephony conditions, as shown in Tab. 4.8. This only requires that exactly those file-based instrumental and subjective results belonging to these grouped telephony conditions are considered for the computations in Eqs. (B.9)–(B.11).

List of Symbols

n, n'	sample indices
ℓ, ℓ_{SPD}	frame indices
k, k', k'_{low}, k'_{high}	frequency bin indices
L, L_{SPD}	numbers of frames
N_{s}, N	frame shift/length (in samples)
N_{-}, N_{+}	frame look-back/-ahead (in samples)
N_{t}	frame transition (in samples)
N_{w}	window length (in samples)
f, Ω	frequencies ($\Omega = 2\pi \frac{f}{f_{\text{s}}}$)
f_{s}, f'_{s}	sampling frequencies
F_0	fundamental frequency
f_{c}, Ω_{c}, $f_{\text{c,p}}$, $f_{\text{c,s}}$	cut-off frequencies
Ω_{M}, g_{M}	modulation frequency/gain
$s_{\text{NB}}(n')$, $s_{\text{LB}}(n)$, $\hat{s}_{\text{UB}}(n)$, $s_{\text{WB}}(n)$	speech data
$w(n)$, $w(n')$	window functions
$S_{\text{NB},\ell}(k')$, $S_{\text{WB},\ell}(k)$	short-term DFT spectra
$\Phi_{\text{NB},\ell}(k')$, $\hat{\Phi}_{\text{UB},\ell}(\tilde{k})$, $\Phi_{\text{WB},\ell}(k)$	short-term DFT power spectra
$\phi_{\text{LB},\ell}(\tilde{n}')$, $\phi_{\text{UB},\ell}(\tilde{n})$, $\hat{\phi}_{\text{WB},\ell}(n)$	ACFs
K_{LB}, K_{UB}, K_{sc}	DFT lengths
k_{NB}, k_{LB}, k_{UB}	mapping functions
$\mathcal{K}_{\text{NB}}$, $\mathcal{K}_{\text{LB}}$, $\mathcal{K}_{\text{UB}}$, $\mathcal{K}_{\text{WB}}$, $\tilde{\mathcal{K}}_{\text{NB}}$, $\tilde{\mathcal{K}}_{\text{LB}}$, $\tilde{\mathcal{K}}_{\text{UB}}$	sets of DFT indices
$\mathbf{a}_{\text{UB},\ell}$, $\mathbf{a}_{\text{WB},\ell}$	LP filter coefficient vectors
$N_{\text{LP}^{(\text{LB})}}$, $N_{\text{LP}^{(\text{UB})}}$, $N_{\text{LP}^{(\text{WB})}}$	LP filter orders
$\hat{\mathbf{r}}_{\text{WB},\ell}$	LP reflection coefficients
$\hat{A}(z)$	z-transform of LPC predictor
$e_{\text{LB}}(n)$, $e_{\text{UB}}(n)$, $\hat{e}_{\text{WB}}(n)$	residual signals
$\sigma_{\text{LB},\ell}$, $\sigma_{\text{UB},\ell}$, $\sigma_{\text{rel},\ell}$	LP prediction gains
g_{UB}, $g_{\text{SPD},\ell}$	static/adaptive residual weighting factor
$\tilde{x}_{\text{acf},\ell}(\,\cdot\,)$, $\tilde{x}_{\text{zcr},\ell}$, $\tilde{x}_{\text{gi},\ell}$, $\tilde{x}_{\text{rfe},\ell}$, $\tilde{x}_{\text{lk},\ell}$, $\tilde{x}_{\text{sc},\ell}$	features

$\tilde{\mathbf{x}}_\ell^{\text{stat}}$, $\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}}$, $\Delta\Delta\tilde{\mathbf{x}}_\ell^{\text{stat}}$, $\tilde{\mathbf{x}}_\ell$, $\mathbf{x}_\ell$	feature vectors
$\tilde{d}$, d	feature dimensions
$g_\ell(\text{n'})$	signal gradient
E_ℓ, $\bar{E}_\ell$, $E_{\text{min},\ell}$, $E_{\text{flr},\ell}$, $\bar{E}_{\text{SPD},\ell}$	frame energies
$\text{sign}(\cdot)$, $\text{rank}(\cdot)$, $E\{\cdot\}$	sign/rank/expectation operators
$\mathbf{H}$	LDA transformation matrix
$\mathbf{B}_{\tilde{\mathbf{x}}}$, $\mathbf{W}_{\tilde{\mathbf{x}}}$	between-/within-class covariance matrix
$\mathbf{J}_{\tilde{\mathbf{x}}}$, $\varsigma_{\tilde{\mathbf{x}}}$, $\varsigma_{\mathbf{x}}(\cdot)$	LDA separability measures
M_{flr}, A_{flr}, $V_{\text{ctr},\ell}$, M_{ctr}, $\Theta_{\text{SPD},\ell}$	SPD parameters
$\mathbf{b}_{\text{LB}}$, N_{LB}	FIR interpolation filter coefficients/order
$\mathbf{b}_{\text{HP}}$	first-order FIR pre-emphasis filter coefficients
α, α_{SPD}, β_{SPD}	first-order IIR smoothing factors
φ_ℓ, $\bar{\varphi}_\ell$, φ'_ℓ	phoneme class labels
ϑ_ℓ, $\bar{\vartheta}_\ell$, ϑ'_ℓ	energy class labels
Φ, Θ	classification thresholds
$\mathcal{P}$	phoneme class alphabet
$N_{\mathcal{P}}$	size of phoneme class alphabet
ρ	positive exponent
$\mathbf{W}^{(1)}$, $\mathbf{b}^{(1)}$, $\mathbf{i}^{(1)}$, $f^{(1)}$, $\mathbf{o}^{(1)}$	ANN parameters of hidden layer
$\mathbf{w}^{(2)}$, $b_1^{(2)}$, $i_1^{(2)}$, $f^{(2)}$, $o_1^{(2)}$	ANN parameters of output layer
s_ℓ, s_ℓ^*	HMM states
λ	set of HMM parameters
$\mathcal{S}$, $\mathcal{S}^{(\varphi)}$	sets of HMM states
$N_{\mathcal{S}}$, $N_{\mathcal{S}}^{(\varphi)}$	numbers of HMM states
$\rho_{j,m}$, $\boldsymbol{\mu}_{j,m}$, $\boldsymbol{\Sigma}_{j,m}$	GMM parameters
M	number of GMM mixture components
$\mathcal{N}(\cdot)$	normal distribution
$\text{P}(\cdot)$, $\text{P}^*(\cdot)$, $\text{P}^{**}(\cdot)$, $\text{P}^{***}(\cdot)$	probability mass functions
$\text{p}(\cdot)$	PDF
π_i	initial state probabilities
$a_{i,j}$	state transition probabilities
$\gamma_\ell(i)$	*a posteriori* probabilities
ε, r	phonetic weighting parameters
$\delta_{\ell+1}(j)$, $\psi_{\ell+1}(j)$	Viterbi score / backtracking pointer
$b_j(\mathbf{x}_\ell)$	state observation likelihood
$\alpha_\ell(i)$, $\beta_\ell(i)$	forward/backward variable of FBA
$\text{H}(\cdot)$, $\text{H}^*(\cdot)$	histogram functions
g, ξ, γ	state transition modification parameters

$\mathcal{C}$	cepstral CB
$\mathbf{c}^{(i)}$	ith CB entry
$\mathbf{c}_\ell$, $\hat{\mathbf{c}}_\ell$	cepstral vectors

List of Abbreviations

ABE	artificial bandwidth extension
ACF	auto-correlation function
ACR	absolute category rating
ADPCM	adaptive differential pulse code modulation
AI	articulation index
AMR	adaptive multi-rate
ANN	artificial neural network
ASR	automatic speech recognition
bagging	bootstrap aggregating
BBC	British Broadcasting Corporation
BCJR	Bahl-Cocke-Jelinek-Raviv
BTE	behind-the-ear
BWE	bandwidth extension
CAN	controller area network
CART	classification and regression tree
CB	codebook
CCR	comparison category rating
CI	confidence interval
CL	confidence level
CMLLR	constrained maximum likelihood linear regression
CMN	cepstral mean normalization
CMOS	comparison mean opinion score
CQS	continuous quality scale
D/A	digital-to-analog
DC	direct current
DCR	degradation category rating
DECT	digital enhanced cordless telecommunications
DET	detection error tradeoff
DFG	German Research Foundation
DFT	discrete Fourier transform

DIAL	diagnostic instrumental assessment of listening quality
EER	equal error rate
EM	expectation maximization
EVS	enhanced voice services
FA	forward algorithm
FAR	false acceptance rate
FBA	forward-backward algorithm
FFT	fast Fourier transform
FIR	finite impulse response
FRR	false rejection rate
GMM	Gaussian mixture model
HD	high definition
HMM	hidden Markov model
HSR	human speech recognition
HTK	Hidden Markov Model Toolkit
IIR	infinite impulse response
IP	internet protocol
IVR	interactive voice response
LB	lower-band
LBG	Linde-Buzo-Gray
LDA	linear discriminant analysis
LDC	Linguistic Data Consortium
LP	linear prediction
LPC	linear predictive coding
LPCC	linear predictive cepstral coefficient
LQO	listening quality objective
LQS	listening quality subjective
LSD	log-spectral distance
MAP	maximum *a posteriori*
MFCC	mel-frequency cepstral coefficient
MIRS	modified intermediate reference system
ML	maximum likelihood
MMSE	minimum mean square error
MNRU	modulated noise reference unit
MOS	mean opinion score
MSE	mean square error
MSIN	mobile station input
MUSHRA	multi-stimulus test with hidden reference and anchor

NB	narrowband
OMA	Oldenburg Measurement Applications
PCM	pulse code modulation
PDF	probability density function
PER	phoneme error rate
PESQ	perceptual evaluation of speech quality
POLQA	perceptual objective listening quality assessment
PSD	power spectral density
RMS	root mean square
RMSE	root mean square error
ROC	receiver operating characteristic
SAMPA	speech assessment methods phonetic alphabet
SAT	speaker adaptive training
SBR	spectral band replication
SCG	scaled conjugate gradient
SF	spectral folding
SLP	selective linear prediction
SNR	signal-to-noise ratio
SPD	speech pause detection
ST	spectral translation
SVM	support vector machine
TOSQA	telecommunication objective speech quality assessment
UB	upper-band
VA	Viterbi algorithm
VAD	voice activity detection
VCV	vowel-consonant-vowel
VoIP	voice over IP
VQ	vector quantization
VTLN	vocal tract length normalization
WB	wideband
WER	word error rate

Bibliography

3GPP S4, "AMR-WB Selection Test Plan (WB-8b) Version 1.0 (Final)," 3rd Generation Partnership Project, Sep. 2000.

3GPP S4/SMG11, "Test Plans for the AMR-WB Qualification Phase (AMR-WB-8a)," 3rd Generation Partnership Project, Mar. 2000.

3GPP SA WG4, "Results of AMR Wideband (AMR-WB) Codec Selection Phase," 3rd Generation Partnership Project, Dec. 2000.

3GPP TS 26.090, "Mandatory Speech Codec Speech Processing Functions; AMR Speech Codec; Transcoding Functions," 3rd Generation Partnership Project, Aug. 1999.

3GPP TS 26.190, "Speech Codec Speech Processing Functions; AMR Wideband Speech Codec; Transcoding Functions," 3rd Generation Partnership Project, Mar. 2001.

3GPP TS 26.290, "Audio Codec Processing Functions; Extended Adaptive Multi-Rate Wideband (AMR-WB+) Codec; Transcoding Functions," 3rd Generation Partnership Project, Sep. 2012.

3GPP TS 26.404, "General Audio Codec Audio Processing Functions; Enhanced aacPlus General Audio Codec; Encoder Specification; Spectral Band Replication (SBR) Part," 3rd Generation Partnership Project, Sep. 2004.

3GPP TS 26.441, "Technical Specification Group Services and System Aspects; Codec for Enhanced Voice Services (EVS); General Overview," 3rd Generation Partnership Project, Dec. 2014.

Y. Agiomyrgiannakis and Y. Stylianou, "Combined Estimation/Coding of Highband Spectral Envelopes for Speech Spectrum Expansion," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Montreal, QC, Canada, May 2004, pp. 469–472.

J. B. Allen, "How Do Humans Process and Recognize Speech?" *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

C. Avendano, H. Hermansky, and E. A. Wan, "Beyond Nyquist: Towards the Recovery of Broad-Bandwidth Speech from Narrow Bandwidth Speech," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, Madrid, Spain, Sep. 1995, pp. 165–168.

L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, Mar. 1974.

P. Bauer and T. Fingscheidt, "An HMM-Based Artificial Bandwidth Extension Evaluated by Cross-Language Training and Test," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, U.S.A., Mar. 2008a, pp. 4589–4592.

P. Bauer and T. Fingscheidt, "Speaker- and Language Dependency of Artificial Bandwidth Extension," in *Proc. of German Annual Conference on Acoustics (DAGA)*, Dresden, Germany, Mar. 2008b, pp. 637–638.

P. Bauer and T. Fingscheidt, "A Statistical Framework for Artificial Bandwidth Extension Exploiting Speech Waveform and Phonetic Transcription," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, Aug. 2009a, pp. 1839–1843.

P. Bauer and T. Fingscheidt, "Spectral Restoration of Narrowband Speech Recordings Supported by Phonetic Transcriptions," in *Proc. of International Conference on Acoustics (NAG/DAGA)*, Rotterdam, The Netherlands, Mar. 2009b, pp. 118–121.

P. Bauer and T. Fingscheidt, "WTIMIT 1.0," Catalog No. LDC2010S02, Linguistic Data Consortium (LDC), Philadelphia, 2010. [Online]. Available: http://catalog.ldc.upenn.edu/LDC2010S02

P. Bauer, T. Fingscheidt, and M. Lieb, "Phonetic Analysis and Redesign Perspectives of Artificial Speech Bandwidth Extension," in *Proc. of Conference on Electronic Speech Signal Processing (ESSV)*, Frankfurt a.M., Germany, Sep. 2008, pp. 215–223.

P. Bauer, M.-A. Jung, and T. Fingscheidt, "Investigations on Offline Artificial Bandwidth Extension of Telephone Speech Databases," in *Proc. of ITG Conference on Speech Communication*, Bochum, Germany, Oct. 2010a.

P. Bauer, M.-A. Jung, J. Qi, and T. Fingscheidt, "On Improving Speech Intelligibility in Automotive Hands-Free Systems," in *Proc. of IEEE International Symposium on Consumer Electronics (ISCE)*, Braunschweig, Germany, Jun. 2010b, pp. 1–5.

P. Bauer, D. Scheler, and T. Fingscheidt, "WTIMIT: The TIMIT Speech Corpus Transmitted Over the 3G AMR Wideband Mobile Network," in *Proc. of ITG Conference on Speech Communication*, Bochum, Germany, Oct. 2010c.

P. Bauer, D. Scheler, and T. Fingscheidt, "WTIMIT: The TIMIT Speech Corpus Transmitted Over The 3G AMR Wideband Mobile Network," in *Proc. of International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010d, pp. 1566–1570.

P. Bauer, R.-L. Fischer, M. Bellanova, H. Puder, and T. Fingscheidt, "On Improving Telephone Speech Intelligibility for Hearing Impaired Persons," in *Proc. of ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 275–278.

P. Bauer, J. Jones, and T. Fingscheidt, "Impact of Hearing Impairment on Fricative Intelligibility for Artificially Bandwidth-Extended Telephone Speech in Noise," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 7039–7043.

P. Bauer, J. Abel, and T. Fingscheidt, "HMM-Based Artificial Bandwidth Extension Supported by Neural Networks," in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes - Juan les Pins, France, Sep. 2014a, pp. 1–5.

P. Bauer, J. Abel, V. Fischer, and T. Fingscheidt, "Automatic Recognition of Wideband Telephone Speech with Limited Amount of Matched Training Data," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, Sep. 2014b, pp. 1232–1236.

P. Bauer, C. Guillaumé, W. Tirry, and T. Fingscheidt, "On Speech Quality Assessment of Artificial Bandwidth Extension," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014c, pp. 6082–6086.

M. Bellanova, M. Serman, M. Latzel, and U. Hoppe, "Entwicklung eines Logatomtests zur Mikroskopischen Differenzierung Unterschiedlicher Hörgerätealgorithmen am Beispiel eines Kompressionsalgorithmus für Hörgeräte," in *Proc. of Annual Conference of Deutsche Gesellschaft für Audiologie (DGA)*, Frankfurt a.M., Germany, Mar. 2010.

M. Bellanova, M. Serman, and M. Latzel, "Non-Adaptive Logatome Testing," Patent IPCOM000205581D, Mar. 2011.

M. Bellanova, M. Mueller-Wehlau, and U. Hoppe, "Subjective Loudness Adjustment: A Method for the Homogenization of German Nonsense-Syllable Speech Material," in *Proc. of Annual Conference of Deutsche Gesellschaft für Audiologie (DGA)*, Erlangen, Germany, Mar. 2012.

L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

H. Carl, "Untersuchung verschiedener Methoden der Sprachkodierung und eine Anwendung zur Bandbreitenvergrößerung von Schmalband-Sprachsignalen," Dissertation, vol. 4 of U. Heute (ed.), Arbeiten über Digitale Signalverarbeitung, Bochum, Germany, 1994.

H. Carl and U. Heute, "Bandwidth Enhancement of Narrow-Band Speech Signals," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Edinburgh, Scotland, Sep. 1994, pp. 1178–1181.

C.-F. Chan and W.-K. Hui, "Quality Enhancement of Narrowband CELP-Coded Speech via Wideband Harmonic Re-Synthesis," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Munich, Germany, Apr. 1997, pp. 1187–1190.

Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical Recovery of Wideband Speech from Narrowband Speech," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Banff, Alberta, Canada, Oct. 1992, pp. 1577–1580.

Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical Recovery of Wideband Speech from Narrowband Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, Oct. 1994.

S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech Enhancement via Frequency Bandwidth Extension Using Line Spectral Frequencies," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Salt Lake City, UT, U.S.A., May 2001, pp. 665–668.

K. W. Church and R. L. Mercer, "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," *Computational Linguistics*, vol. 19, no. 1, pp. 1–24, Mar. 1993.

M. H. Cohen, J. P. Giangola, and J. Balogh, *Voice User Interface Design*. Addison-Wesley, 2004.

M. Cooke and O. Scharenborg, "The Interspeech 2008 Consonant Challenge," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brisbane, Australia, Sep. 2008, pp. 1765–1768.

M. G. Croll, "Sound-Quality Improvement of Broadcast Telephone Calls," Technical Report 1972/26, The British Broadcasting Corporation (BBC), Jan. 1972.

N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Springer, 2011.

A. Das and J. H. L. Hansen, "Constrained Iterative Speech Enhancement Using Phonetic Classes," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1869–1883, Aug. 2012.

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

M. Dietrich, "Performance and Implementation of a Robust ADPCM Algorithm for Wideband Speech Coding with 64 kbit/s," in *Proc. of International Zürich Seminar on Digital Communications*, Zürich, Switzerland, Mar. 1984, pp. 15–21.

N. Enbom and W. B. Kleijn, "Bandwidth Expansion of Speech Based on Vector Quantization of the Mel Frequency Cepstral Coefficients," in *Proc. of IEEE Workshop on Speech Coding (SCW)*, Porvoo, Finland, Jun. 1999, pp. 171–173.

Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

J. Epps and W. H. Holmes, "A New Technique for Wideband Enhancement of Coded Narrowband Speech," in *Proc. of IEEE Workshop on Speech Coding (SCW)*, Porvoo, Finland, Jun. 1999, pp. 174–176.

H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed., ser. Springer Series in Information Sciences. Springer, 2007, vol. 22.

E. T. Federighi, "Extended Tables of the Percentage Points of Student's t-distribution," *Journal of the American Statistical Association*, vol. 54, no. 287, pp. 683–688, Sep. 1959.

S. Ferraz de Campos Neto and K. Järvinen, "Wideband Speech Coding Standards and Wireless Services: Guest Editorial," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 56–57, May 2006.

T. Fingscheidt, "The Silent Speech Bandwidth Revolution in Mobile Telephony," IEEE Speech and Language Processing Technical Committee (SLTC) Newsletter, Aug. 2012. [Online]. Available: http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2012-08/SpeechBandwidthRevolutionInMobileTelephony/

T. Fingscheidt, "Spoken Language Processing Lecture Notes," Summer Term 2014, Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany.

T. Fingscheidt and P. Bauer, "A Phonetic Reference Paradigm for Instrumental Speech Quality Assessment of Artificial Speech Bandwidth Extension," in *Proc. of International Workshop on Perceptual Quality of Systems (PQS)*, Vienna, Austria, Sep. 2013, pp. 36–39.

M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe-Verbmobil Speech Recognition Engine," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Munich, Germany, Apr. 1997, pp. 83–86.

V. Fischer and S. Kunzmann, "The EML Transcription Platform - A Flexible Transcription Environment for Robust Speech Recognition," in *Proc. of International Conference on Acoustics (AIA/DAGA)*, Merano, Italy, Mar. 2013, pp. 2063–2066.

J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. Springer, 1972.

H. Fletcher and R. H. Galt, "The Perception of Speech and Its Relation to Telephony," *Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 89–151, Mar. 1950.

B. Fodor and T. Fingscheidt, "Reference-free SNR Measurement for Narrowband and Wideband Speech Signals in Car Noise," in *Proc. of ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 199–202.

R. B. Frary, "Formula Scoring of Multiple-Choice Tests (Correction for Guessing)," *Educational Measurement: Issues and Practice*, vol. 7, no. 2, pp. 33–38, Jun. 1988.

N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the Regeneration of Wideband Speech from Narrowband Speech," *Journal on Applied Signal Processing*, vol. 2001, no. 1, pp. 266–274, Jan. 2001.

K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990.

J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Catalog No. LDC93S1, Linguistic Data Consortium (LDC), Philadelphia, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1

B. Geiser, P. Jax, and P. Vary, "Artificial Bandwidth Extension of Speech Supported by Watermark-Transmitted Side Information," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, Lisbon, Portugal, Sep. 2005a, pp. 1497–1500.

B. Geiser, P. Jax, and P. Vary, "Robust Wideband Enhancement of Speech by Combined Coding and Artificial Bandwidth Extension," in *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, Sep. 2005b, pp. 21–24.

B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaume, and S. Ragot, "Bandwidth Extension for Hierarchical Speech and Audio Coding in ITU-T Rec. G.729.1," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2496–2509, Nov. 2007.

Global Mobile Suppliers Association, "Mobile HD Voice: Global Update Report," May 2016. [Online]. Available: http://gsacom.com/paper/mobile-hd-voice-global-update-report-2/

A. Graves, A. r. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.

M. T. Hagan, H. B. Demuth, and M. Beale, *Neural Network Design*. PWS Publishing Company, 1996.

J. Han, G. Mysore, and B. Pardo, "Language Informed Bandwidth Expansion," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sep. 2012, pp. 1–6.

J. H. L. Hansen and B. L. Pellom, "Text-Directed Speech Enhancement Employing Phone Class Parsing and Feature Map Constrained Vector Quantization," *Speech Communication*, vol. 21, no. 3, pp. 169–189, Mar. 1997.

E. Hänsler and G. Schmidt, *Speech and Audio Processing in Adverse Environments*. Springer, 2008, vol. 1.

H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

M. Hosoki, T. Nagai, and A. Kurematsu, "Speech Signal Band Width Extension and Noise Removal Using Subband HMM," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Orlando, FL, U.S.A., May 2002, pp. 245–248.

X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Prentice Hall, 2001.

G. W. Hughes and M. Halle, "Spectral Properties of Fricative Consonants," *Journal of the Acoustical Society of America*, vol. 28, no. 2, pp. 303–310, Mar. 1956.

B. Iser and G. Schmidt, "Neural Networks versus Codebooks in an Application for Bandwidth Extension of Speech Signals," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 565–568.

B. Iser, W. Minker, and G. Schmidt, *Bandwidth Extension of Speech Signals.* Springer, 2008, vol. 13.

ITU-R BS.1534-1, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," International Telecommunication Union, Jan. 2003.

ITU-T G.191, "Software Tool Library 2009 User's Manual," International Telecommunication Union, Nov. 2009.

ITU-T G.711.1, "Wideband Embedded Extension for G.711 Pulse Code Modulation," International Telecommunication Union, Mar. 2008.

ITU-T G.722, "7 kHz Audio-Coding within 64 kbit/s," International Telecommunication Union, Nov. 1988.

ITU-T G.722.1, "Coding at 24 and 32 kbit/s for Hands-Free Operation in Systems with Low Frame Loss," International Telecommunication Union, Sep. 1999.

ITU-T G.729EV, "Quality Assessment Characterisation / Optimisation Step1 Test Plan for the ITU-T G.729 Based Embedded Variable Bit-Rate (G.729EV) Extension to the ITU-T G.729 Speech Codec," International Telecommunication Union, Nov. 2005.

ITU-T P.1401, "Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models," International Telecommunication Union, Jul. 2012.

ITU-T P.341, "Transmission Characteristics for Wideband Digital Loudspeaking and Hands-Free Telephony Terminals," International Telecommunication Union, Mar. 2011.

ITU-T P.56, "Objective Measurement of Active Speech Level," International Telecommunication Union, Dec. 2011.

ITU-T P.800, "Methods for Subjective Determination of Transmission Quality," International Telecommunication Union, Aug. 1996.

ITU-T P.800.1, "Mean Opinion Score (MOS) Terminology," International Telecommunication Union, Jul. 2006.

ITU-T P.810, "Modulated Noise Reference Unit (MNRU)," International Telecommunication Union, Feb. 1996.

ITU-T P.830, "Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs," International Telecommunication Union, Feb. 1996.

ITU-T P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," International Telecommunication Union, Feb. 2001.

ITU-T P.862.1, "Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO," International Telecommunication Union, Nov. 2003.

ITU-T P.862.2, "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," International Telecommunication Union, Nov. 2007.

ITU-T P.863, "Perceptual Objective Listening Quality Assessment," International Telecommunication Union, Jan. 2011.

ITU-T P.863.1, "Application Guide for Recommendation ITU-T P.863," International Telecommunication Union, May 2013.

V. Iyengar, R. Rabipour, P. Mermelstein, and B. R. Shelton, "Speech Bandwidth Extension Method and Apparatus," Patent US 5,455,888 A, Oct. 1995. [Online]. Available: http://www.google.com/patents/US5455888

J. Abel, "HMM-basierte Erweiterung der akustischen Bandbreite von Sprachdatenbanken," Master's Thesis, Institut für Nachrichtentechnik, Technische Universität Braunschweig, Braunschweig, Germany, Nov. 2013.

J. Jones, "Design and Optimization of a Phonetic Classifier to Support Artificial Bandwidth Extension in Real-Time Applications," Master's Thesis, Institut für Nachrichtentechnik, Technische Universität Braunschweig, Braunschweig, Germany, Jan. 2013.

P. Jax, "Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds," Dissertation, vol. 15 of P. Vary (ed.), Aachener Beiträge zu digitalen Nachrichtensystemen, Aachen, Germany, 2002.

P. Jax and P. Vary, "Wideband Extension of Telephone Speech Using a Hidden Markov Model," in *Proc. of IEEE Workshop on Speech Coding (SCW)*, Delavan, WI, U.S.A., Sep. 2000, pp. 133–135.

P. Jax and P. Vary, "Artificial Bandwidth Extension of Speech Signals Using MMSE Estimation Based on a Hidden Markov Model," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Hong Kong, China, Apr. 2003, pp. 680–683.

K. Kalgaonkar, "Probabilistic Space Maps for Speech with Applications," Dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, U.S.A., 2011.

K. Kalgaonkar and M. Clements, "Vocal Tract Area Based Artificial Bandwidth Extension," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Cancun, Mexico, Oct. 2008, pp. 480–485.

K. Kalgaonkar and M. A. Clements, "Sparse Probabilistic State Mapping and its Application to Speech Bandwidth Expansion," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 4005–4008.

M. Karafiat, L. Burget, J. Cernocky, and T. Hain, "Application of CMLLR in Narrow Band Wide Band Adapted Systems," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium, Aug. 2007, pp. 282–285.

I. Katsir, I. Cohen, and D. Malah, "Speech Bandwidth Extension Based on Speech Phonetic Content and Speaker Vocal Tract Shape Estimation," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Aug. 2011, pp. 461–465.

I. Katsir, D. Malah, and I. Cohen, "Evaluation of a Speech Bandwidth Extension Algorithm Based on Vocal Tract Shape Estimation," in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012, pp. 1–4.

G. Keidser, H. Dillon, M. Flax, T. Ching, and S. Brewer, "The NAL-NL2 Prescription Procedure," *Audiology Research*, vol. 1, no. 1, pp. 88–90, 2011.

L. J. Kepler, M. Terry, and R. H. Sweetman, "Telephone Usage in the Hearing-Impaired Population," *Ear and Hearing*, vol. 13, no. 5, pp. 311–319, Oct. 1992.

J. Kontio, L. Laaksonen, and P. Alku, "Neural Network-Based Artificial Bandwidth Expansion of Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 873–881, Mar. 2007.

U. Kornagel, "Techniques for Artificial Bandwidth Extension of Telephone Speech," *Signal Processing*, vol. 86, no. 6, pp. 1296–1306, Jun. 2006.

W. Krebber, "Sprachübertragungsqualität von Fernsprech-Handapparaten," Dissertation, VDI-Fortschrittsberichte, Reihe 10, Nr. 357, Aachen, Germany, 1995.

K. R. Krishnamachari, R. E. Yantorno, and J. M. Lovekin, "Use of Local Kurtosis Measure for Spotting Usable Speech Segments in Co-Channel Speech," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Salt Lake City, UT, U.S.A., May 2001, pp. 649–652.

L. Laaksonen and J. Virolainen, "Binaural Artificial Bandwidth Extension (B-ABE) for Speech," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 4009–4012.

L. Laaksonen, J. Kontio, and P. Alku, "Artificial Bandwidth Expansion Method to Improve Intelligibility and Quality of AMR-Coded Narrowband Speech," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Philadelphia, PA, U.S.A., Mar. 2005, pp. 809–812.

L. Laaksonen, V. Myllylä, and R. Niemistö, "Evaluating Artificial Bandwidth Extension by Conversational Tests in Car Using Mobile Devices with Integrated Hands-Free Functionality," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy, Aug. 2011, pp. 1177–1180.

E. Larsen, R. M. Aarts, and M. Danessis, "Efficient High-Frequency Bandwidth Extension of Music and Speech," in *112th Convention of the Audio Engineering Society (AES)*, Munich, Germany, May 2002, pp. 1–5.

K.-F. Lee and H.-W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

F. Li and J. B. Allen, "Manipulation of Consonants in Natural Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 496–504, Mar. 2011.

Y.-F. Liao, J.-S. Lin, and W.-H. Tsai, "Bandwidth Mismatch Compensation for Robust Speech Recognition," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland, Sep. 2003, pp. 3093–3096.

Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney, "The RWTH 2007 TC-STAR Evaluation System for European English and

Spanish," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium, Aug. 2007, pp. 2145–2148.

C. Lopes and F. Perdigão, "Phone Recognition on the TIMIT Database," Speech Technologies, Chapter 14, p. 679, Jun. 2011. [Online]. Available: http://www.intechopen.com/books/speech-technologies/phoneme-recognition-on-the-timit-database

T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *Journal on Applied Signal Processing*, vol. 2005, no. 1, pp. 1110–1126, Jan. 2005.

S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration Strategies for Audio-Visual Speech Processing: Applied to Text-Dependent Speaker Recognition," *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 495–506, Jun. 2005.

D. Macho and Y. M. Cheng, "On the Use of Wideband Signal for Noise Robust ASR," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Hong Kong, Apr. 2003, pp. 109–112.

J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

J. Makhoul and M. Berouti, "High-Frequency Regeneration in Speech Coding Systems," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, DC, U.S.A., Apr. 1979, pp. 428–431.

T. Marill, "Automatic Recognition of Speech," *IRE Transactions on Human Factors in Electronics*, vol. HFE-2, no. 1, pp. 34–38, Mar. 1961.

J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, 1st ed. Springer, 1976, vol. 12.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 4, Rhodes, Greece, Sep. 1997, pp. 1899–1903.

R. Martin, "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, Berlin, Germany, Sep. 1993, pp. 1093–1096.

R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

S. Möller, M. Wältermann, B. Lewcio, N. Kirschnick, and P. Vidales, "Speech Quality While Roaming in Next Generation Networks," in *Proc. of IEEE International Conference on Communications (ICC)*, Dresden, Germany, Jun. 2009, pp. 1–5.

S. Möller, E. Kelaidi, F. Köster, N. Côté, P. Bauer, T. Fingscheidt, T. Schlien, H. Pulakka, and P. Alku, "Speech Quality Prediction for Artificial Bandwidth Extension Algorithms," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, Aug. 2013, pp. 3439–3443.

S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publishers, 2000.

A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, J. Allen, and S. Euler, "SpeechDat-Car: A Large Speech Database for Automotive Environments," in *Proc. of International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, May 2000, pp. 895–900.

C. Nadeu and M. Tolos, "Recognition Experiments with the SpeechDat-Car Aurora Spanish Database Using 8 kHz- and 16 kHz-Sampled Signals," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Italy, Dec. 2001, pp. 135–138.

Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of Broadband Speech From Narrowband Speech Using Piecewise Linear Mapping," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 3, Rhodes, Greece, Sep. 1997, pp. 1643–1646.

M. Nilsson and W. B. Kleijn, "Avoiding Over-Estimation in Bandwidth Extension of Telephony Speech," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Salt Lake City, UT, U.S.A., May 2001, pp. 869–872.

M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn, "Gaussian Mixture Model Based Mutual Information Estimation Between Frequency Bands in Speech," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Orlando, FL, U.S.A., May 2002, pp. 525–528.

D. Nolden, H. Ney, and R. Schlüter, "Time Conditioned Search in Automatic Speech Recognition Reconsidered," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, Sep. 2010, pp. 234–237.

A. H. Nour-Eldin and P. Kabal, "Combining Frontend-Based Memory with MFCC Features for Bandwidth Extension of Narrowband Speech," in *Proc. of IEEE International Confer-*

ence on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, Apr. 2009, pp. 4001–4004.

NTT-AT, "Multi-Lingual Speech Database for Telephonometry 1994," NTT Advanced Technology Corporation, 1994. [Online]. Available: http://www.ntt-at.com/product/speech/

D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, Aug. 1999.

A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Prentice Hall, 1989.

P. Bauer, "Artificial Bandwidth Extension with Multilingual Training Process," Diploma Thesis, Institut für Nachrichtentechnik, Technische Universität Braunschweig, Braunschweig, Germany, Jul. 2007.

A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. McGraw-Hill, 2002.

K.-Y. Park and H. S. Kim, "Narrowband to Wideband Conversion of Speech Using GMM Based Transformation," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Istanbul, Turkey, Jun. 2000, pp. 1843–1846.

P. J. Patrick, "Enhancement of Bandlimited Speech Signal," Dissertation, Loughborough University of Technology, Loughborough, Great Britain, 1983.

E. Paulus, *Sprachsignalverarbeitung: Analyse, Erkennung, Synthese*, 1st ed. Spektrum Akademischer Verlag, 1998.

B. Pfister and T. Kaufmann, *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Springer, 2008.

T. V. Pham, F. Schaefer, and G. Kubin, "A Novel Implementation of the Spectral Shaping Approach for Artificial Bandwidth Extension," in *Proc. of International Conference on Communications and Electronics (ICCE)*, Nha Trang, Vietnam, Aug. 2010, pp. 262–267.

J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 4th ed. Pearson Prentice Hall, 2007.

H. Pulakka, "Development and Evaluation of Artificial Bandwidth Extension Methods for Narrowband Telephone Speech," Dissertation, School of Electrical Engineering, Aalto University, Helsinki, Finland, 2013.

H. Pulakka and P. Alku, "Bandwidth Extension of Telephone Speech Using a Neural Network and a Filter Bank Implementation for Highband Mel Spectrum," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, Sep. 2011.

H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an Artificial Speech Bandwidth Extension Method in Three Languages," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1124–1137, Aug. 2008.

H. Pulakka, U. Remes, K. Palomaki, M. Kurimo, and P. Alku, "Speech Bandwidth Extension Using Gaussian Mixture Model-Based Estimation of the Highband Mel Spectrum," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5100–5103.

H. Pulakka, L. Laaksonen, V. Myllylä, S. Yrttiaho, and P. Alku, "Conversational Evaluation of Speech Bandwidth Extension Using a Mobile Handset," *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 203–206, Feb. 2012a.

H. Pulakka, L. Laaksonen, V. Myllylä, S. Yrttiaho, and P. Alku, "Conversational Evaluation of Artificial Bandwidth Extension of Telephone Speech Using a Mobile Handset," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012b, pp. 4069–4072.

H. Pulakka, L. Laaksonen, S. Yrttiaho, V. Myllylä, and P. Alku, "Conversational Quality Evaluation of Artificial Bandwidth Extension of Telephone Speech," *Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 848–861, Aug. 2012c.

H. Pulakka, U. Remes, S. Yrttiaho, K. Palomäki, M. Kurimo, and P. Alku, "Bandwidth Extension of Telephone Speech to Low Frequencies Using Sinusoidal Synthesis and a Gaussian Mixture Model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2219–2231, Oct. 2012d.

H. Pulakka, A. Rämö, V. Myllylä, H. Toukomaa, and P. Alku, "Subjective Voice Quality Evaluation of Artificial Bandwidth Extension: Comparing Different Audio Bandwidths and Speech Codecs," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, Sep. 2014, pp. 2804–2808.

L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

T. Ramabadran and M. Jasiuk, "Artificial Bandwidth Extension of Narrow-Band Speech Signals via High-Band Energy Estimation," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008, pp. 1–5.

M. Reif, F. Shafait, M. Goldstein, T. Breuel, and A. Dengel, "Automatic Classifier Selection for Non-Experts," *Pattern Analysis and Applications*, vol. 17, no. 1, pp. 83–96, Feb. 2014.

D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

S. M. Ross, *Statistik für Ingenieure und Naturwissenschaftler*, 3rd ed. Elsevier, 2006.

D. Rybach, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, P. Lehnen, J. Lööf, D. Nolden, C. Plahl, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RWTH ASR - The RWTH Aachen University Speech Recognition System," RWTH Aachen University, 2001. [Online]. Available: http://www-i6.informatik.rwth-aachen.de/rwth-asr/

D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The RWTH Aachen University Open Source Speech Recognition System," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, U.K., Sep. 2009, pp. 2111–2114.

D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, U.S.A., Dec. 2011, pp. 1–4.

S. Walz, "Iterative Decodierung in der Spracherkennung," Diploma Thesis, Institut für Nachrichtentechnik, Technische Universität Braunschweig, Braunschweig, Germany, Sep. 2011.

M. Sanna and M. Murroni, "A Codebook Design Method for Fricative Enhancement in Artificial Bandwidth Extension," in *Proc. of International ICST Mobile Multimedia Communications Conference (MOBIMEDIA)*, London, UK, Sep. 2009, pp. 39:1–39:7.

P. Scalart and J. V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Atlanta, GA, U.S.A., May 1996, pp. 629–632.

R. E. Schapire, "The Strength of Weak Learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, Jun. 1990.

K.-O. Schmidt, "Neubildung von unterdrückten Sprachfrequenzen durch ein nichtlinear verzerrendes Glied," *Telegraphen- und Fernsprech-Technik*, vol. 22, no. 1, pp. 13–22, Jan. 1933.

K. Schnell and A. Lacroix, "Time-Varying Linear Prediction for Speech Analysis and Synthesis," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, U.S.A., Mar. 2008, pp. 3941–3944.

M. L. Seltzer and A. Acero, "Training Wideband Acoustic Models Using Mixed-Bandwidth Training Data via Feature Bandwidth Extension," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Philadelphia, PA, U.S.A., Mar. 2005, pp. 921–924.

M. L. Seltzer and A. Acero, "Training Wideband Acoustic Models Using Mixed-Bandwidth Training Data for Speech Recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 1, pp. 235–245, Jan. 2007.

M. L. Seltzer, A. Acero, and J. Droppo, "Robust Bandwidth Extension of Noise-Corrupted Narrowband Speech," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, Lisbon, Portugal, Sep. 2005, pp. 1509–1512.

P. Setiawan, "Exploration and Optimization of Noise Reduction Algorithms for Speech Recognition in Embedded Devices," Dissertation, Berichte aus der Informationstechnik, Universität der Bundeswehr München, München, Germany, 2009.

C. Shahnaz, W. P. Zhu, and M. . Ahmad, "A Robust Pitch Estimation Algorithm in Noise," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Honolulu, HI, U.S.A., Apr. 2007, pp. 1073–1076.

A. J. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

Suhadi, "Speech Enhancement Using Data-Driven Concepts," Dissertation, Band 24 von T. Fingscheidt (Ed.), Mitteilungen aus dem Institut für Nachrichtentechnik der Technischen Universität Braunschweig, Braunschweig, Germany, 2012.

Y. Sunil and R. Sinha, "Exploration of Class Specific ABWE for Robust Children's ASR under Mismatched Condition," in *Proc. of International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, Jul. 2012, pp. 1–5.

Y. Tanaka and N. Hatazoe, "Reconstruction of Wideband Speech From Telephone-Band Speech by Multilayer Neural Networks," Spring Meeting of ASJ, 1-4-19, pp. 255-256, 1995.

E. Terhardt, *Akustische Kommunikation*. Springer, 1998.

M. R. P. Thomas, J. Gudnason, P. A. Naylor, B. Geiser, and P. Vary, "Voice Source Estimation for Artificial Bandwidth Extension of Telephone Speech," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, U.S.A., Mar. 2010, pp. 4794–4797.

A. Uncini, F. Gobbi, and F. Piazza, "Frequency Recovery of Narrow-Band speech Using Adaptive Spline Neural Networks," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Phoenix, AZ, U.S.A., Mar. 1999, pp. 997–1000.

T. Unno and A. McCree, "A Robust Narrowband to Wideband Extension System Featuring Enhanced Codebook Mapping," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Philadelphia, PA, U.S.A., Mar. 2005, pp. 805–808.

V. Välimäki, "Discrete-Time Modeling of Acoustic Tubes Using Fractional Delay Filters," Dissertation, School of Electrical Engineering, Aalto University, Helsinki, Finland, 1995.

V. Välimäki and T. I. Laakso, "Suppression of Transients in Time-Varying Recursive Filters for Audio Signals," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 6, Seattle, WA, U.S.A., May 1998, pp. 3569–3572.

P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, 2006.

A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 2000.

L. Welling, N. Haberland, and H. Ney, "Acoustic Front-End Optimization for Large Vocabulary Speech Recognition," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, Sep. 1997, pp. 2099–2102.

L. Welling, S. Kanthak, and H. Ney, "Improved Methods for Vocal Tract Normalization," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Phoenix, AZ, U.S.A., Mar. 1999, pp. 761–764.

C. Weng, D. Yu, S. Watanabe, and B.-H. Juang, "Recurrent Deep Neural Networks for Robust Speech Recognition," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 5532–5536.

R. Wiese, *Phonetik und Phonologie*, 1st ed. UTB, 2010.

C. Yagli, M. A. T. Turan, and E. Erzin, "Artificial Bandwidth Extension of Spectral Envelope Along a Viterbi Path," *Speech Communication*, vol. 55, no. 1, pp. 111–118, Jan. 2013.

S. Yao and C.-F. Chan, "Block-Based Speech Bandwidth Extension System with Separated Envelope Energy Ratio Estimation," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, Sep. 2005, pp. 1–4.

S. Yao and C.-F. Chan, "Speech Bandwidth Enhancement Using State Space Speech Dynamics," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Toulouse, France, May 2006, pp. 489–492.

H. Yasukawa, "Enhancement of Telephone Speech Quality by Simple Spectrum Extrapolation Method," in *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 2, Madrid, Spain, Sep. 1995, pp. 1545–1548.

Y. Yoshida and M. Abe, "An Algorithm to Reconstruct Wideband Speech from Narrowband Speech Based on Codebook Mapping," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan, Sep. 1994, pp. 1591–1594.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.4)," University of Cambridge, Cambridge, United Kingdom, Dec. 2006. [Online]. Available: http://htk.eng.cam.ac.uk/

G. Yu, W. Russell, R. Schwartz, and J. Makhoul, "Discriminant Analysis and Supervised Vector Quantization for Continuous Speech Recognition," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Albuquerque, NM, U.S.A., Apr. 1990, pp. 685–688.

Own Publications

C. Voges, **P. Bauer**, and T. Fingscheidt, "A Particle Filtering Algorithm for Audiovisual Speaker Localisation," in *Proc. of Workshop on Positioning, Navigation and Communication (WPNC)*, Hannover, Germany, Mar. 2007, pp. 103–108.

P. Bauer and T. Fingscheidt, "Bessere Qualität der Telefonie mit künstlicher Erweiterung der Sprachbandbreite," *Fachzeitschrift für Informations- und Kommunikationstechnik (ntz), VDE Verlag*, vol. 2, no. 1, pp. 2–5, Feb. 2008.

P. Bauer and T. Fingscheidt, "An HMM-Based Artificial Bandwidth Extension Evaluated by Cross-Language Training and Test," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, U.S.A., Mar. 2008, pp. 4589–4592.

P. Bauer and T. Fingscheidt, "Speaker- and Language Dependency of Artificial Bandwidth Extension," in *Proc. of German Annual Conference on Acoustics (DAGA)*, Dresden, Germany, Mar. 2008, pp. 637–638.

P. Bauer, T. Fingscheidt, and M. Lieb, "Phonetic Analysis and Redesign Perspectives of Artificial Speech Bandwidth Extension," in *Proc. of Conference on Electronic Speech Signal Processing (ESSV)*, Frankfurt a.M., Germany, Sep. 2008, pp. 215–223.

P. Bauer and T. Fingscheidt, "A Statistical Framework for Artificial Bandwidth Extension Exploiting Speech Waveform and Phonetic Transcription," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, Aug. 2009, pp. 1839–1843.

P. Bauer and T. Fingscheidt, "Spectral Restoration of Narrowband Speech Recordings Supported by Phonetic Transcriptions," in *Proc. of International Conference on Acoustics (NAG/DAGA)*, Rotterdam, The Netherlands, Mar. 2009, pp. 118–121.

P. Bauer and T. Fingscheidt, "WTIMIT 1.0," Linguistic Data Consortium (LDC), Philadelphia, Mar. 2010. [Online]. Available: http://catalog.ldc.upenn.edu/LDC2010S02

P. Bauer, M.-A. Jung, and T. Fingscheidt, "Investigations on Offline Artificial Bandwidth

Extension of Telephone Speech Databases," in *Proc. of ITG Conference on Speech Communication*, Bochum, Germany, Oct. 2010.

P. Bauer, M.-A. Jung, J. Qi, and T. Fingscheidt, "On Improving Speech Intelligibility in Automotive Hands-Free Systems," in *Proc. of IEEE International Symposium on Consumer Electronics (ISCE)*, Braunschweig, Germany, Jun. 2010, pp. 1–5.

P. Bauer, D. Scheler, and T. Fingscheidt, "WTIMIT: The TIMIT Speech Corpus Transmitted Over the 3G AMR Wideband Mobile Network," in *Proc. of ITG Conference on Speech Communication*, Bochum, Germany, Oct. 2010.

P. Bauer, D. Scheler, and T. Fingscheidt, "WTIMIT: The TIMIT Speech Corpus Transmitted Over The 3G AMR Wideband Mobile Network," in *Proc. of Conference on International Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010, pp. 1566–1570.

P. Bauer, R.-L. Fischer, M. Bellanova, H. Puder, and T. Fingscheidt, "On Improving Telephone Speech Intelligibility for Hearing Impaired Persons," in *Proc. of ITG Conference on Speech Communication*, Braunschweig, Germany, Sep. 2012, pp. 275–278.

P. Bauer, J. Jones, and T. Fingscheidt, "Impact of Hearing Impairment on Fricative Intelligibility for Artificially Bandwidth-Extended Telephone Speech in Noise," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 7039–7043.

T. Fingscheidt and **P. Bauer**, "A Phonetic Reference Paradigm for Instrumental Speech Quality Assessment of Artificial Speech Bandwidth Extension," in *Proc. of International Workshop on Perceptual Quality of Systems (PQS)*, Vienna, Austria, Sep. 2013, pp. 36–39.

S. Möller, E. Kelaidi, F. Köster, N. Côté, **P. Bauer**, T. Fingscheidt, T. Schlien, H. Pulakka, and P. Alku, "Speech Quality Prediction for Artificial Bandwidth Extension Algorithms," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, Aug. 2013, pp. 3439–3443.

P. Bauer, J. Abel, and T. Fingscheidt, "HMM-Based Artificial Bandwidth Extension Supported by Neural Networks," in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes - Juan les Pins, France, Sep. 2014, pp. 1–5.

P. Bauer, J. Abel, V. Fischer, and T. Fingscheidt, "Automatic Recognition of Wideband Telephone Speech with Limited Amount of Matched Training Data," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, Sep. 2014, pp. 1232–1236.

P. Bauer, C. Guillaumé, W. Tirry, and T. Fingscheidt, "On Speech Quality Assessment of Artificial Bandwidth Extension," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 6082–6086.