Department of Signal Processing and Acoustics

Artificial bandwidth extension of narrowband speech - enhanced speech quality and intelligibility in mobile devices

Laura Laaksonen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 64/2013

Artificial bandwidth extension of narrowband speech - enhanced speech quality and intelligibility in mobile devices

Laura Laaksonen

A doctoral dissertation completed for the degree of Doctor of Science in Technology to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held in the lecture hall S1 of the school on May 3, 2013, at 12 noon.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics

Supervising professor
Professor Paavo Alku

Preliminary examiners
Professor Gernot Kubin, Graz University of Technology, Austria
Professor Yannis Stylianou, University of Crete, Greece

Opponent
Professor Bayya Yegnanarayana, International Institute of Information Technology (IIIT), Hyderabad, India

Aalto University publication series
DOCTORAL DISSERTATIONS 64/2013

ISBN 978-952-60-5124-6 (printed)
ISBN 978-952-60-5125-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5125-3

Unigrafia Oy
Helsinki 2013

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Laura Laaksonen

Name of the doctoral dissertation
Artificial bandwidth extension of narrowband speech - enhanced speech quality and intelligibility in mobile devices

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 64/2013

Field of research Acoustics and audio signal processing

Manuscript submitted 27 September 2012 **Date of the defence** 3 May 2013

Permission to publish granted (date) 10 January 2013 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Even today, most of the telephone users are offered only narrowband speech transmission. The limited frequency band from 300 Hz to 3400 Hz reduces both quality and intelligibility of speech due to the missing high frequency components that are important cues especially in consonant sounds. Particularly in mobile communications that often takes place in noisy environments, degraded speech intelligibility results in listener fatigue and difficulty in speaker recognition. The deployment of wideband (50–7000 Hz), and superwideband (50–140000 Hz) speech transmission is ongoing, but the current narrowband speech coding will coexist with the new technologies still for years.

In this thesis, a speech enhancement method called artificial bandwidth extension (ABE) for narrowband speech is studied. ABE methods aim to improve quality and intelligibility of narrowband speech by regenerating the missing high frequency content in the speech signal, typically in the frequency range 4 kHz–8 kHz. Since the enhanced speech quality is achieved without any transmitted information, the algorithm can be implemented at the receiving end of a communication link, for example in a mobile device after decoding the speech signal.

This thesis presents algorithms for artificially extending the speech bandwidth. The methods are primarily designed for monaural speech signals, but also the extension of binaural speech signals is addressed. The algorithms are developed such that they incur reasonable computational costs, memory consumption, and algorithmic delays for mobile communications. These and other implementational issues related to mobile devices are addressed here.

The performance of the methods has been evaluated by several subjective tests, including listening-opinion tests in several languages, intelligibility tests, and conversational tests. The evaluations have been mostly carried out with coded speech to provide realistic results. The results from the subjective evaluations of the methods show that artificial bandwidth extension can improve quality and intelligibility of narrowband speech signals in mobile communications. Further evidence of the reliability of the methods has been obtained by successful product implementations.

Keywords speech processing, speech enhancement, artificial bandwidth extension, speech quality, mobile device

ISBN (printed) 978-952-60-5124-6 **ISBN (pdf)** 978-952-60-5125-3

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2013

Pages 170 **urn** http://urn.fi/URN:ISBN:978-952-60-5125-3

Tekijä
Laura Laaksonen

Väitöskirjan nimi
Puheen keinotekoinen kaistanlaajennus - parempilaatuista ja ymmärrettävämpää puhetta matkapuhelimiin

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 64/2013

Tutkimusala Akustiikka ja äänenkäsittelytekniikka

Käsikirjoituksen pvm 27.09.2012 **Väitöspäivä** 03.05.2013

Julkaisuluvan myöntämispäivä 10.01.2013 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Suurin osa puhelinliikenteestä on vielä nykyäänkin kapeakaistaista, eli puhelignaalista lähetetään vain 300–3400 Hz:in taajuuskaista. Rajoitettu taajuuskaista huonontaa sekä puheen laatua että ymmärrettävyyttä, koska korkeataajuiset, erityisesti konsonanttiäänteille tärkeät akustiset vihjeet, puuttuvat signaalista. Etenkin meluisissa ympäristöissä matkapuhelimien puhesignaalien heikko ymmärrettävyys väsyttää käyttäjiä sekä aiheuttaa ongelmia puhujan tunnistettavuudessa. Vaikka laajakaistaisen (50–7000 Hz) puheensiirtotekniikan käyttöönotto on aloitettu, kapeakaistaiset puheensiirtomenetelmät ovat käytössä vielä vuosia uusien menetelmien rinnalla.

Tässä väitöskirjassa tutkitaan kapeakaistaisen puhesignaalin keinotekoista kaistanlaajennusta. Tällä puheenparannusmenetelmällä pyritään parantamaan puheäänen laatua ja ymmärrettävyyttä lisäämällä puhesignaaliin sisältöä puuttuville taajuuksille, esimerkiksi 4–8 kHz:in taajuuskaistalle. Koska puuttuvan kaistan alkuperäisestä sisällöstä ei lähetetä mitään tietoa, laajennus voidaan toteuttaa puhelinyhteyden vastaanottopäässä, kuten vastaanottajan matkapuhelimessa puhesignaalin dekoodauksen jälkeen.

Tässä työssä esitellään keinotekoisia kaistanlaajennusalgoritmeja. Algoritmit on suunniteltu ensisijaisesti monosignaaleille, mutta myös binauraalisen signaalin laajennusta on tutkittu. Algoritmikehityksessä huomioitiin matkapuhelinympäristön asettamat laskenta-, muisti- sekä algoritmiviiverajoitukset. Näitä ja muita menetelmään liittyviä tuotteistusasioita on myös käsitelty tässä tutkimuksessa.

Kaistanlaajennusmenetelmien laatua on mitattu useilla subjektiivisiilla testeillä, kuten eri kielillä toteutetuilla kuuntelukokeilla ja keskustelukokeilla. Näissä laadunarvioinneissa on käytetty pääasiassa koodattua puhemateriaalia, jotta tulokset olisivat mahdollisimman todenmukaisia. Laadunarviointitulokset osoittavat, että keinotekoisella kaistanlaajennuksella pystytään parantamaan kapeakaistaisen puheen laatua ja ymmärrettävyyttä matkapuhelinympäristössä. Tätä tulosta tukevat myös algoritmin onnistuneet matkapuhelintoteutukset.

Avainsanat puheenkäsittely, puheen siistaus, keinotekoinen kaistanlaajennus, puheen laatu, matkapuhelin

ISBN (painettu) 978-952-60-5124-6 **ISBN (pdf)** 978-952-60-5125-3

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2013

Sivumäärä 170 **urn** http://urn.fi/URN:ISBN:978-952-60-5125-3

Preface

This thesis is a story of a long term research collaboration project between Aalto University, Department of Signal Processing and Acoustics, and Nokia. The collaboration in the field of artificial bandwidth extension (ABE) of telephone speech signals, started in 1999, and I have had the opportunity to be part of it since 2002. The story started from an idea of an ABE algorithm. During the years, several ABE algorithms were developed and evaluated by subjective tests. The next step was to implement them in mobile products, and to evaluate the algorithms in realistic conversational context. I have enjoyed this project and learned so much about speech processing technology, scientific research work, and speech quality, thanks to many supporting and delightful people from both the university and Nokia.

First of all, I would like to thank Prof. Paavo Alku for giving me the opportunity to work on ABE in a first place. He hired me to the Laboratory of Acoustics and Audio Signal Processing to work on ABE and to write a M.Sc. thesis. In 2003 I started my career in Nokia and the same year, I came up with an idea to start the PhD studies. Fortunately, Paavo welcomed me to be one of his PhD students. Paavo, without your support and encouragement this thesis would not have been finished. I appreciate how you always find time to review, comment, and discuss research work, no matter how busy you are.

Another important person behind the ABE collaboration project is Jari Sjöberg, the leader of the Audio Algorithms team in Nokia. Thank you for letting me concentrate on the ABE research, and writing of this thesis. I look forward to returning back to work in September.

All the publications of my thesis have been written together with talented people and I wish to thank all my co-authors. From Juho Kontio I learned a lot about neural networks. Thanks to Hannu Pulakka, it has been a pleasure working with you during these years. In addition, thanks to Martti Vainio, Jouni Pohjalainen and Santeri Yrttiaho for your help and participation in the publications.

I'm grateful to the pre-examiners, Prof. Gernot Kubin and Prof. Yannis Stylianou for their dedicated work and valuable comments on the

manuscript. I would also like to thank Luis Costa for proof-reading the manuscript.

Furthermore, I would like to express my gratitude to the Nokia Audio Algorithms team in Tampere. Many of my former and current colleagues have influenced this work. Päivi Valve, from you I learned that it is possible to find a solution for every problem, one way or another. Ville Myllylä, I appreciate our discussions, and your innovative ideas. Riitta Niemistö, I remember you once said, that you believe I can finish my dissertation some day. I have kept that in my mind during the times when I wasn't quite sure myself. Jukka Vartiainen, thanks for your help in many signal processing questions. In addition, I would like to thank Matti Kajala, Erkki Paajanen, Antti Pasanen, Anu Lahdenpohja, Jouko Salo, and Eero Niemelä for your participation in ABE work. I also wish to thank Anssi Rämö and Henri Toukomaa for listening test arrangements.

In Nokia Helsinki site, I have met many great audio people. I would like to thank you for inspiring lunch breaks, parties, and discussions during the years. Especially, I would like to thank my friends Riitta Väänänen, Julia Turku, Jussi Virolainen and Jarmo Hiipakka.

I would like to thank many acoustics people from Otaniemi. Conference trips to Philadelphia, Lisbon, and Florence were so much fun because of you. I remember fun dinners in Philadelphia. In Lisbon, the fado concert was amazing. And the bus trip across the Europe in the middle of the night was unforgettable. It's no wonder I don't really remember the discussions on my poster after travelling (and being 7 months pregnant) from Frankfurt to Florence by bus because the flights to Italy were cancelled.

Finally, I would like to express my gratitude to my family and friends. Thank you Anna for your friendship. Miika, thank you for everything, for being so supportive and positive during this project, and for photoshooting the cover picture for my thesis on a freezing cold winter day. Ellen (6 years), Lotta (4 years) and Lauri (1 year), you are the world to me. I would also like to thank my parents, Elina and Jorma, for all the encouragement and support. Thanks to my sister Eeva, and my brother Eero, and their families for all the good times and laughs. And last but not least, thanks to Merja and Pekka for all the help. Finalizing the thesis while staying at home with three children would not have been possible without the help from my whole family.

Espoo, March 22, 2013,

Laura Laaksonen

Contents

List of publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Laura Laaksonen, Juho Kontio, and Paavo Alku. Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, volume 1, pages 809–812, March 2005.

II Juho Kontio, Laura Laaksonen, and Paavo Alku. Neural network-based artificial bandwidth expansion of speech. *IEEE Transactions on Audio, Speech, and Language Processing*, volume 15, issue 3, pages 873–881, March 2007.

III Hannu Pulakka, Laura Laaksonen, Martti Vainio, Jouni Pohjalainen, and Paavo Alku. Evaluation of an artificial speech bandwidth extension method in three languages. *IEEE Transactions on Audio, Speech, and Language Processing*, volume 16, issue 6, pages 1124–1137, August 2008.

IV Laura Laaksonen, Hannu Pulakka, Ville Myllylä, and Paavo Alku. Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal. *IEEE Transactions on Consumer Electronics*, volume 55, issue 2, pages 780–787, May 2009.

V Laura Laaksonen and Jussi Virolainen. Binaural artificial bandwidth extension (B-ABE) for speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, volume 1, pages 4009–4012, April 2009.

VI Laura Laaksonen, Ville Myllylä, and Riitta Niemistö. Evaluating artificial bandwidth extension by conversational tests in car using mobile devices with integrated hands-free functionality. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech*, pages 1177–1180, August 2011.

VII Hannu Pulakka, Laura Laaksonen, Santeri Yrttiaho, Ville Myllylä, and Paavo Alku. Conversational quality evaluation of artificial bandwidth extension of telephone speech. *The Journal of the Acoustical Society of America*, volume 132, issue 2, pages 848–861, August 2012.

Author's contribution

Publication I: "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech"

The author was the main developer of the algorithm. The author planned the quality and intelligibility tests together with the co-authors. She also processed the samples for the tests. The author primarily wrote the article.

Publication II: "Neural network-based artificial bandwidth expansion of speech"

The author planned the subjective test and sample processing together with the co-authors. She performed the objective analysis and wrote a considerable part of the article. The author also participated in the development of the algorithm.

Publication III: "Evaluation of an artificial speech bandwidth extension method in three languages"

The author was the main developer of the algorithm. She planned the subjective listening test together with the co-authors and processed the speech samples for the test. The author wrote the algorithm description in section II for the article.

Publication IV: "Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal"

The author was the main developer of the algorithm. She had a significant role in planning and organizing the listening tests. In addition, the author primarily wrote the article except for some parts in sections III.C and IV.B.

Publication V: "Binaural artificial bandwidth extension (B-ABE) for speech"

The author developed the algorithm and planned the subjective tests together with the co-author. She conducted the listening tests and analysed the results. The author wrote most of the article.

Publication VI: "Evaluating artificial bandwidth extension by conversational tests in car using mobile devices with integrated hands-free functionality"

The author planned, designed, and conducted the conversational tests with the co-authors. She primarily wrote the article except for section 2.1.

Publication VII: "Conversational quality evaluation of artificial bandwidth extension of telephone speech"

The author planned and designed the conversational tests together with the co-authors. She also participated in piloting the tests. She is the main developer of the method ABE1 and implemented the algorithm for the test. She wrote sections I, II A, and parts of III for the article.

List of abbreviations

2G	second generation
3G	third generation
3GPP	3rd Generation Partnership Project
ABE	artificial bandwidth extension
ACELP	algebraic code excited linear prediction
ACR	absolute category rating
ADPCM	adaptive differential pulse code modulation
AMPS	Advanced Mobile Phone System
AMR	adaptive multirate
AMR-WB	adaptive multirate wideband
B-ABE	binaural artificial bandwidth extension
CCR	comparison category rating
CEPT	European Conference of Postal and Telecommunications Administrations
CMOS	comparison mean opinion score
CS-ACELP	conjugate structure algebraic code-excited linear prediction
DCR	degradation category rating
DFT	discrete Fourier transform
DMOS	degradation mean opinion score

DRT	diagnostic rhyme test
EFR	enhanced full rate
EM	expectation maximization
ETSI	European Telecommunications Standards Institute
FFNN	feedforward neural network
FFT	fast Fourier transform
GMM	Gaussian mixture model
GSM	Global System for Mobile Communications
HD	high definition
HMM	hidden Markov model
HRTF	head related transfer function
ILD	interaural level difference
IMS	internet protocol multimedia system
ITD	interaural time difference
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
LD-CELP	low-delay code excited linear prediction
log-PCM	logarithmic pulse code modulation
LP	linear prediction
LPC	linear predictive coding
LSD	log spectral distortion
LSF	line spectrum frequency,
LSP	line spectrum pair
LTE	Long Term Evolution
MFCC	mel frequency cepstral coefficient
MI	mutual information
MIPS	millions of instructions per second

MLP	multilayer perceptron
MMSE	minimum mean square error
MOS	mean opinion score
MRT	modified rhyme test
NEABE	neuroevolution artificial bandwidth extension
NEAT	neuroevolution of augmenting topologies
NMT	Nordic Mobile Telephone
PCM	pulse code modulation
PDF	probability density function
PESQ	perceptual evaluation of speech quality
POLQA	perceptual objective listening quality analysis
PSTN	public switched telephone network
RPE-LTP	regular pulse excitation with long term prediction
SD	spectral distortion
SNR	signal-to-noise ratio
SRT	speech reception threshold
TFO	tandem-free operation
TrFO	transcoder-free operation
UMTS	Universal Mobile Telecommunications System
VoIP	voice over internet protocol
VoLTE	voice over Long Term Evolution
VSELP	vector sum excited linear prediction
WCDMA	wideband code division multiple access

List of figures

1. Introduction

Speech signals in telephone communications have been bandlimited since the beginning of the history of the telephone. During the days of analogue telephone, the limited bandwidth was due to the physical restrictions of acoustic components and the bandwidth capacity. A digital transmission utilizing pulse code modulation (PCM) adopted a 8 kHz sampling rate and the speech bandwidth of 300–3400 Hz both for compatibility with the analogue telephone and also for reasons of bandwidth capacity [1]. For decades, consumers were offered only narrowband (also called telephone band) speech transmission. The telephone users got used to the *telephone speech* that sounds muffled and has reduced speech quality [2] and intelligibility [3], especially during consonants, due to the missing important high frequency acoustic cues.

The narrowband PCM quality may have been adequate for landline telephony in the 20$^{\text{th}}$ century, but mobile communications has brought new challenges and demands for speech transmission. In the 1990's, the number of mobile phones started to increase rapidly. The first coders designed for the 2G mobile networks suffered from degraded speech quality compared to the narrowband PCM. Later, the enhanced full rate (EFR) [4] and adaptive multirate (AMR) [5] codecs reached nearly the narrowband PCM quality. The first significant improvement to the speech quality and intelligibility was achieved by increasing the speech bandwidth and the sampling rate of a speech codec. The adaptive multirate wideband (AMR-WB) speech codec with a frequency band of 50–7000 Hz was standardized in 2001 [6], and its deployment started in 2009 [7]. Still today, only a small portion of end users in mobile telephony are offered wideband transmission that is marketed as high definition (HD) voice. The upgrade of networks and mobile devices for AMR-WB support is time consuming. On the other hand, in voice over internet protocol (VoIP) applications wideband

or superwideband speech is often supported, for example in [8].

Speech transmission in mobile networks is characterized by the fact that mobile phones can be used everywhere. Background noise conditions may vary from quiet to extremely noisy and complex acoustic surroundings. From a speech quality perspective, the small mechanical components, i.e. earpieces and microphones, as well as the variety of possible bluetooth, car and other accessories that are used with the mobile devices are also a challenge. To face these challenges, a proper acoustical design of the device and speech enhancement methods that modify the speech signal at both ends of the telephone link are needed. Speech enhancement algorithms aim to improve the quality and intelligibility of speech, for example, by reducing noise and echo from the signal or by emphasizing perceptually important parts of the signal. Noise cancellation and single-channel and multichannel noise reduction techniques are examples of such algorithms that are important in mobile communications. These methods are applied in modern mobile devices and networks.

Motivated by the slow deployment of wideband speech, one of the speech enhancement research topics since the mid 1990's has been artificial bandwidth extension (ABE). Narrowband speech transmission and coding uses a sampling rate of 8 kHz that restricts the speech bandwidth to 300–3400 Hz. ABE methods aim to improve quality and intelligibility by regenerating the missing high frequency content of a speech signal in the receiving end of the transmission. An ABE method increases the sampling rate, typically from 8 kHz to 16 kHz, and adds new frequency components to the highband, i.e. typically a frequency range of 4–8 kHz. The extension is completely artificial, indicating that no information related to the missing highband is transmitted. However, there also are methods that are not completely artificial but utilize transmitted side information in the extension procedure [9, 10, 11].

ABE can be seen as a speech enhancement method for narrowband speech signals that improves quality and intelligibility. Especially in noisy environments, the wider bandwidth is beneficial. On the other hand, ABE can be seen as an algorithm that transforms a narrowband signal to wideband in the receiving terminal when wideband transmission is not available. During the transition phase from narrowband to wideband speech transmission, the speech bandwidth may vary between and even during phone calls, depending on the available network conditions and telephone devices. The challenge in ABE methods is to generate as natural sound-

ing wideband speech as possible. While pursuing this goal, some unnatural highband artefacts might be created in the signal that might annoy listeners. As for all speech enhancement methods, artificial bandwidth extension should be as transparent to end users as possible.

Subjective speech quality assessment is extremely important in the field of ABE research. Both underestimation and overestimation of the signal level in the artificial highband is likely to produce an audible artefact. Especially fricatives and plosives, which are characterized by a burst of frication, are extremely sensitive to unnatural signal components in the highband, because they have a considerable amount of energy in frequencies above 4 kHz. For implementing an ABE method in a mobile device, thorough testing and evaluation of the method is needed. The interoperability over the whole signal path, including other speech enhancements and acoustical properties of the device, has to be taken into account in the implementation of the ABE feature.

1.1 Aim of the study

This thesis studies artificial bandwidth extension methods for narrowband speech signals. Most of the research work has been carried out within a collaboration research project between Nokia and the Department of Signal Processing and Acoustics of the Aalto University. The project on artificial bandwidth extension started in 1999, but the work related to this thesis has been conducted during the years 2003–2011 in Nokia. The thesis addresses four main research topics:

1. Development of ABE algorithms for narrowband telephone speech that are robust with respect to noisy and distorted (due to speech coding) input signals

2. ABE of binaural speech signals

3. Comprehensive evaluation of ABE methods by subjective listening-only tests and conversational tests

4. Implementation of an ABE method in a mobile device

The algorithm development includes new algorithms that improve speech

quality and intelligibility of narrowband telephone speech. These methods are presented in Publication I, Publication II, Publication III and Publication VII. Furthermore, the extension of binaural signals is addressed in Publication V. The designed methods are evaluated by several listening opinion tests, including, for example, listening tests addressing potential language dependency of the algorithm (Publication III) and conversational tests (Publication VI and Publication VII). Finally, the implementation of an ABE method in a mobile device is discussed in Publication IV.

2. Speech and hearing

2.1 Speech production

In everyday life we say that we speak with our mouths. In fact, the mouth is needed for speech communication, but by itself it is not at all adequate for speech production. Actually, speech production starts from the lungs, where the lung pressure is increased by drawing air into the lungs. While maintaining some pressure in the lungs, air is forced from the lungs and passed through the trachea, larynx, and the vocal tract. The organs involved in speech production are shown in the schematic illustration of figure 2.1.

The larynx is a sophisticated organ that controls the air flow from the lungs. In the larynx, two horizontal ligaments called vocal folds are attached posteriorly to the arytenoid cartilages that, in turn, are used to control the size of the opening of the vocal folds. This opening is called the *glottis*.

The vocal tract is an acoustic tube that starts from the larynx, ends at the lips, and consists of the pharyngeal cavity and the oral cavity. The total length of the vocal tract from the larynx to the lips is about 17 cm for an adult male and 13.5 cm for a female [12]. By changing the length and the cross section profiles of the vocal tract, mostly by moving the lips, jaw, tongue and velum, humans are able to produce different speech sounds. The influence of the vocal tract on the speech sound is called *articulation*. The velum separates an ancillary path, the nasal tract, from the vocal tract for sound transmission. It starts from the velum and ends at the nostrils. During nasal speech sounds this path is opened, whereas during non-nasal sounds the velum is tightly drawn up.

There are three main ways how humans use the speech production or-

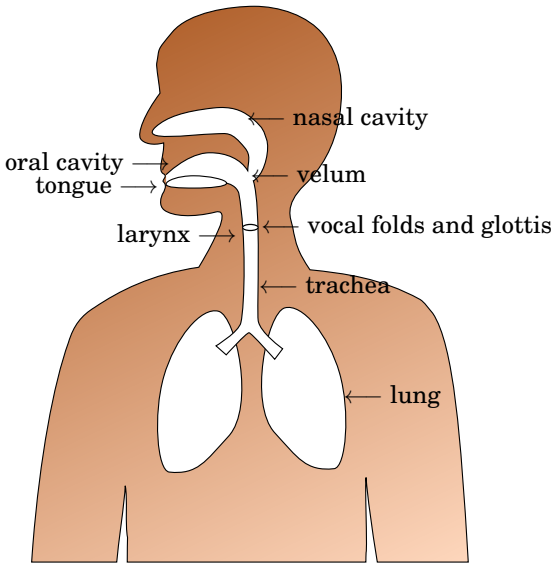

Figure 2.1. Organs involved in speech production.

gans to produce different speech sounds [13]. These three ways result in the speech sound categories: 1) voiced sounds, 2) unvoiced sounds, and 3) plosives (strictly speaking, in the classification of [13], plosives can be either voiced or unvoiced). Voiced sounds refer to quasi-periodic sounds, such as vowels and nasals, during which the air from the lungs travels through the larynx, where two vocal folds start to open due to the increased air pressure. After a complete opening, the vocal folds start to close due to the Bernoulli effect until they are completely closed. The resulting quasi-periodic signal is the source signal for voiced sounds and is called a glottal pulse. The production of a sound in this manner is called *phonation* [14].

A simplified glottal pulse waveform is shown in figure 2.2. The periodicity originates from the vibrating vocal folds that open and close regularly. The vocal folds are completely open at the maximum amplitude of the glottal pulse waveform and closed at the minimum. The round shape of the glottal pulse can be explained by the watery tissue of the vocal folds, and this round waveform shape results in low-pass characteristics in the frequency domain. The fundamental frequency, f_0, and consequently the perceived pitch of speech is determined by the vibration rate of the vocal folds. For females, f_0 is typically about 200 Hz and for males, 120 Hz. In the frequency domain, the spectrum of the glottal pulse has a comb-

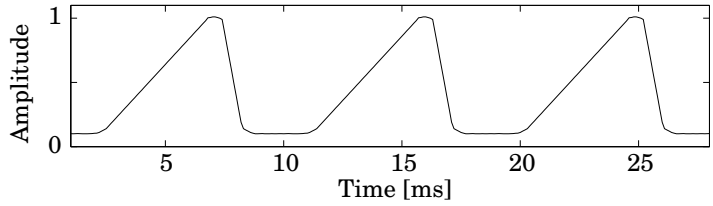

Figure 2.2. Simplified glottal pulse waveform.

shaped structure that shows the fundamental frequency and its harmonics.

During the production of voiced sounds the vocal tract is either completely or partly open. The vocal tract acts as an acoustic filter that creates resonances called *formants*. Since the parameters of the vocal tract as an acoustical filter are continuous and distributed over the entire tract, the resulting transfer function depends on the overall length, shape, and volume of the vocal tract rather than just a single parameter.

During unvoiced sounds that refer to noise-like sounds, such as fricatives, the vocal folds are almost completely open, and no periodic glottal pulse is created. Instead, the source signal is noise generated by a turbulent air flow through a constriction in the vocal tract. The constriction is formed, for example, by the tongue behind the teeth. The noise source is further modified by the resonances of the vocal tract and radiated from the mouth [14].

During plosives, the vocal tract is completely closed at some place, for example by the lips, and the air flow is blocked. When the obstruction is suddenly opened, the released airflow from the lungs produces a sudden impulse in pressure causing a short, audible sound with a noise-like waveform.

2.2 Signal characteristics of speech sounds

2.2.1 Voiced sounds

Western languages comprise mostly voiced sounds. For example, about 78 % of speech sounds in standard English have been reported to be voiced [14]. A typical voiced speech sound is shown in figure 2.3 both as a time-domain waveform and as a frequency-domain spectrum. The waveform is characterized by a periodic structure and a large variation in amplitude.

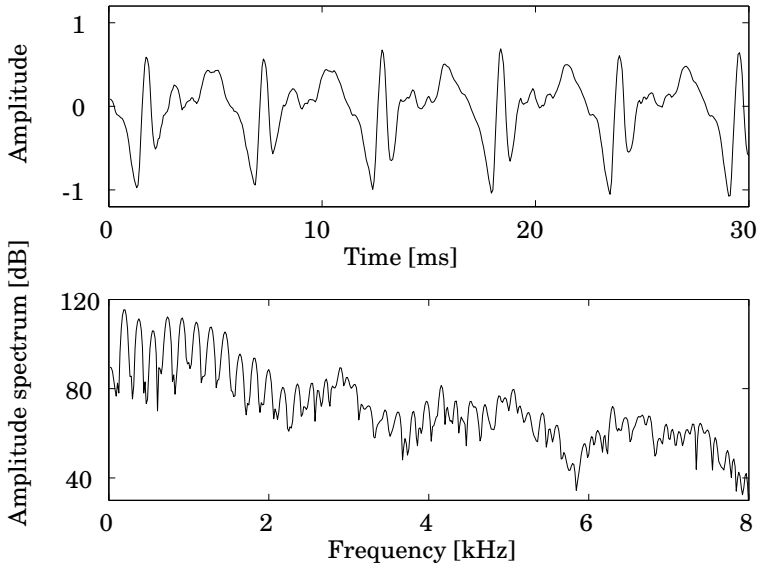

Figure 2.3. Typical voiced speech sound (Finnish [a]) presented as a time domain waveform (top), and its amplitude spectrum (bottom). The spectrum has been computed with a 1024-point FFT using a Hann window.

The amplitude spectrum shows the clear harmonic structure, especially at low frequencies. The first harmonic corresponds to the fundamental frequency. The formant frequencies can also be identified from the maximum peaks of the spectral envelope. The spectrum of voiced sounds typically has low-pass characteristics, which originates from the excitation signal, i.e. the glottal pulse.

In a narrowband signal, due to the low-pass characteristics, a great deal of the energy of a voiced sound is preserved despite the restricted bandwidth. In addition, a narrowband signal also contains the most important harmonics. Although the fundamental frequency may be missing, the human ear is still able to hear the pitch correctly, a phenomenon called the *missing fundamental*. From the bandwidth extension point of view, the most important aim is to extend the low-pass envelope in the higher frequencies. This low-pass characteristics originating from the glottal pulse is also one of the justifications for the correlation between the narrowband speech signal and missing high frequencies. On the other hand, the exact regeneration of the harmonic structure at higher frequencies is not as perceptually important [15].

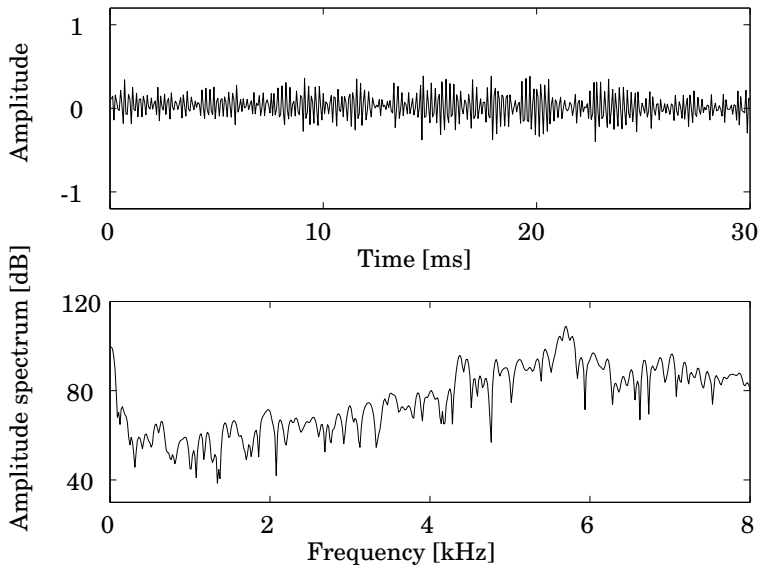

Figure 2.4. Typical unvoiced speech sound (Finnish [s]) presented as a time domain waveform (top), and its amplitude spectrum (bottom). The spectrum has been computed with a 1024-point FFT using a Hann window.

2.2.2 Unvoiced sounds

A waveform of a typical unvoiced speech sound is shown in figure 2.4. The small amplitude values and rapid changes of direction in the temporal signal waveform are due to the noisy source signal. The amplitude spectrum in the lower part of figure 2.4 increases with frequency, indicating that unvoiced sounds are characterized by high frequency components. It is evident that a large portion of the energy of fricative sounds is missing from narrowband signals. These speech sounds are especially challenging for ABE methods, since natural sounding wideband fricatives are obtained only if an adequate amount of energy is added to the higher frequencies. On the other hand, misplaced strong frequency components in high frequencies are likely to result in severe artefacts.

2.2.3 Plosives

A waveform and an amplitude spectrum of a typical plosive sound is presented in figure 2.5. Plosives are characterized by a short silent period caused by a break in voicing, followed by a short burst of frication as the pressure in the place of constriction is suddenly released in the vocal

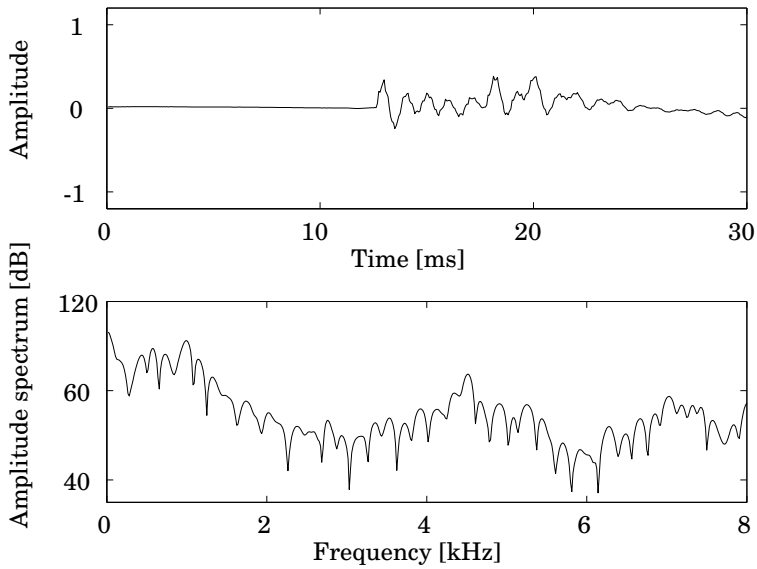

Figure 2.5. Typical plosive (Finnish [k]) presented as a time domain waveform (top), and its amplitude spectrum (bottom). The spectrum has been computed with a 1024-point FFT using a Hann window.

tract. After the burst comes a voicing period that leads to the following vowel. In addition to unvoiced sounds, plosives are also challenging for ABE methods. If the amplitudes of the added high frequency components of a plosive are too large, it is easily perceived as a tingle.

2.3 Sounds of speech

Even though speech as an acoustic signal is a continuous waveform, in a linguistic domain it comprises a finite number of discrete distinguishable sounds called *phonemes*. A person who knows a certain language identifies these phonemes because they have a distinctive linguistic function, although the acoustic properties of signals representing a certain phoneme are both speaker and context dependent. A classification of speech sounds often starts from a division into vowels and consonants. Then the consonants and vowels are further classified according to the manner and place of articulation. For example, the Finnish vowels are /a, e, i, o, u, y, ä, ö/. The Finnish consonants are classified as plosives /k, p, t, g, b, d/, fricatives /f, s, h/, nasals /n, m, η /, trills /r/, laterals /l/, and semivowels /j, v/.

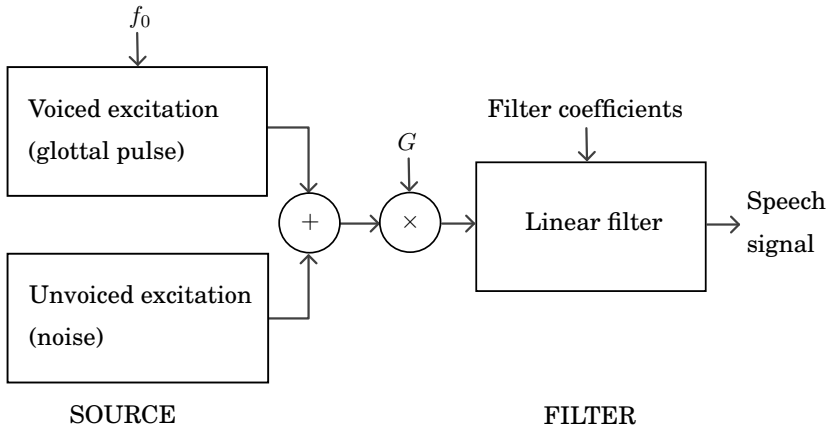

Figure 2.6. Block diagram of a source filter model of speech production. The fundamental frequency, f_0 , can be given as a parameter for the voiced excitation. The gain control, G, is for controlling the energy level of the signal.

2.4 Source filter model

Speech production can be modelled with a source-filter model [16]. According to the model, the two parts, namely the source model and the filter model, are independent. Although the assumption is not completely justified, since there is some interaction between the glottal source and the filter, the model usually yields adequate results. The model consists of two alternative sources, a quasi-periodic pulse generator modelling the glottal pulse for voiced sounds and a noise signal modelling a constriction in the vocal tract for unvoiced sounds, as shown in figure 2.6. The fundamental frequency can be given as a parameter to the pulse generator. The gain control, G, is needed to control the energy level of the signal. The vocal tract and the nasal tract are modelled independently of the source signal by a linear time-varying filter.

2.4.1 Linear prediction

The source-filter model has been applied to many areas of speech processing, for instance in speech analysis, speech synthesis, and speech coding. In speech coding, so called vocoders utilize the source filter model to reduce the number of parameters needed to characterize speech sounds. The parametrization of the source and the filter separately, instead of the whole waveform, has been an effective way to reduce the bit rate in speech coding. The vocal tract filter and the excitation signal can be estimated

from a speech signal using a well known technique called linear prediction (LP), or linear predictive coding (LPC) [17].

LP is based on the simple idea that the next signal sample, $s(n)$, can be estimated as a linear combination of the p previous samples:

$$\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k). \tag{2.1}$$

The parameters $a(k)$ are unknown, and they are solved by minimizing the mean square of the energy of the error signal, $e(n)$, between the real sample $s(n)$ and the estimate $\hat{s}(n)$. The error signal, also called the residual, can be written as

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a(k)s(n-k). \tag{2.2}$$

As a result from an autocorrelation method, the optimal LPC prediction coefficients $\mathbf{A} = \begin{pmatrix} a(1) & a(2) & \dots & a(p) \end{pmatrix}^T$ are obtained from

$$\mathbf{A} = \mathbf{R^{-1}} \cdot \mathbf{R}', \tag{2.3}$$

where $\mathbf{R}$ is an autocorrelation matrix of the form

$$\mathbf{R} = \begin{pmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ \vdots & \vdots & \ddots & \ddots \\ R(p-1) & R(p-2) & \dots & R(0) \end{pmatrix} \tag{2.4}$$

and $\mathbf{R}'$ is an autocorrelation vector

$$\mathbf{R}' = \begin{pmatrix} R(1) & R(2) & \dots & R(p) \end{pmatrix}^T. \tag{2.5}$$

LPC analysis and synthesis

The LPC analysis parametrizes an input speech signal with a residual signal, p LPC coefficients, and a gain factor, G. First, the order of LPC, i.e. the parameter p, the frame size, and the window function have to be defined. Typical window sizes are about 10–30 ms, a period of time where features such as mean power, frequency spectrum, and probability density distribution may be considered to remain relatively constant [12]. As an example, a 20-ms frame of the Finnish speech sound [a], windowed with a rectangular and a Hann window, are shown in figure 2.7.

For the windowed speech frame, the autocorrelation parameters, $a(k)$, are computed from equation 2.3. The residual signal shown in figure 2.8

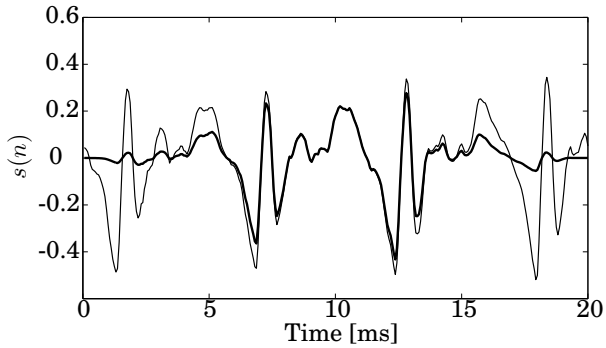

Figure 2.7. 20-ms frame of vowel [a], windowed with a rectangular (thin line) and a Hann window (bold line).

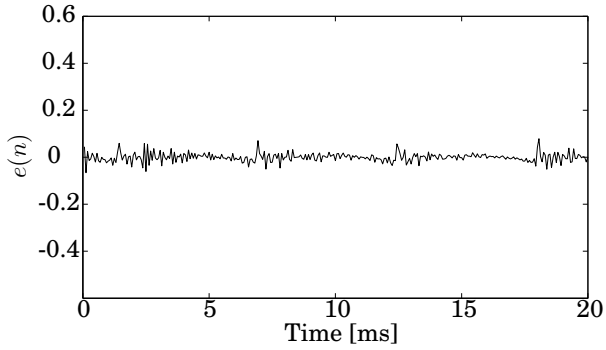

Figure 2.8. LPC residual signal, $e(n)$, for vowel [a] calculated from the signal in figure 2.7.

is obtained by filtering the original un-windowed frame with the obtained inverse filter of the form

$$A(z) = 1 - \sum_{k=1}^{p} a(k) z^{-k}. \tag{2.6}$$

Finally, the gain factor, G, is calculated from the original frame.

On the synthesis side, the original speech signal is reconstructed by filtering the residual signal with an LPC synthesis filter of the form

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a(k) z^{-k}}. \tag{2.7}$$

The LPC synthesis filter models the spectral envelope of the signal. In practice, the order of the LPC, p defines how accurately the filter models the overall spectral shape, the formants, and the fine structure of the signal. Therefore, the parameter p is also directly proportional to the

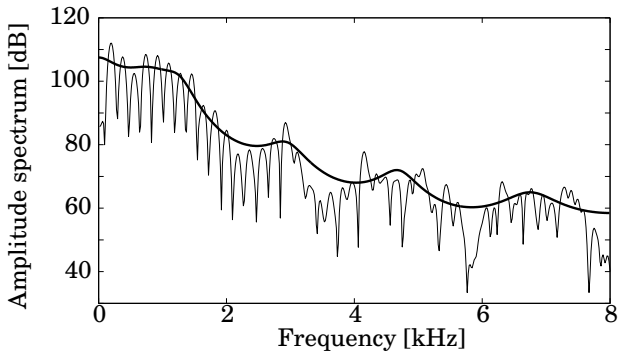

Figure 2.9. Amplitude spectrum of vowel [a] (thin line) and an LPC spectrum of order 12 (bold line).

prediction gain that is often used to measure the performance of LPC: a higher LPC order yields a more accurate LPC model. Figure 2.9 shows the amplitude spectrum of vowel [a] and the corresponding LPC spectrum of order 12.

2.5 Hearing

The human ear, shown schematically in figure 2.10, can be divided into the outer ear, the middle ear, and the inner ear. The outer ear consists of the pinna and the ear canal which terminates at the eardrum. The pinna protects the ear canal but also facilitates the localization of sound sources [14]. The ear canal acts as an acoustic tube having its first resonance at about 4 kHz. The eardrum transforms the pressure variations of incoming sound waves into mechanical vibrations [18].

In the middle ear, the mechanical vibrations are transmitted to the inner ear by three bones, the ossicles (the malleus, the incus and the stapes). The ossicles perform an impedance transformation between the air medium of the outer ear and the liquid of the inner ear. At the oval window, where the middle ear ends, the pressure is about 30 times the pressure at the eardrum. Between the middle ear and the oral cavity is the Eustachian tube that equalizes the pressure between the middle ear and the outer ear during, for example, swallowing or during air pressure changes due to rapid change in altitude.

In the inner ear, the spiral cochlea begins at the oval window, and it transforms the vibrations into properly coded neural impulses. Vibrations arriving at the cochlea make the basilar membrane vibrate. Next to

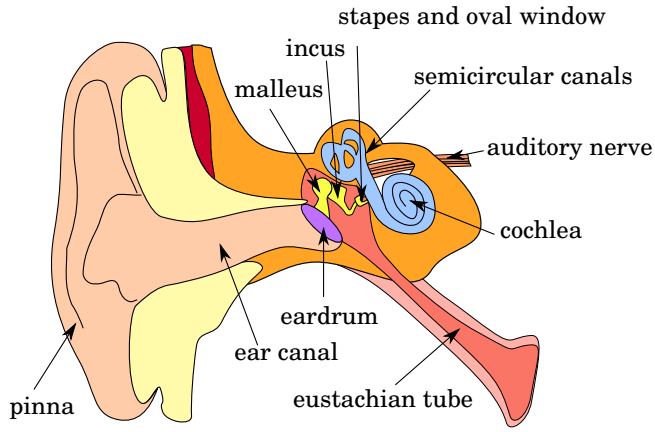

Figure 2.10. Schematic illustration of the human ear.

the basilar membrane, is the organ of Corti that contains hair cells, approximately 20000–30000 hair cells in several rows. The movement of the basilar membrane activates the hair cells that, in turn, excite neurons in the auditory nerve.

2.5.1 Binaural hearing and localization

Communication systems have traditionally used *monaural* hearing, where sound is perceived only by one ear or by two ears with no difference between the signals heard by the two ears. However, when sounds are received by two ears, the human hearing system also analyses the spatial and localization information in the signals. This *binaural* hearing can be beneficial in many situations, like noisy, complex auditory environments. An example of the benefits of binaural hearing is the well-known effect called the "cocktail party effect". It refers to the fact that human listeners can concentrate on listening to one speaker when others are talking simultaneously or when background noise is present [19].

From the speech transmission point of view spatial audio could be beneficial, for example, in teleconferencing systems. The participants of a teleconference might be virtually placed at different positions around the listener. Since the performance of 3D audio is dependent on the bandwidth of the signal, ABE methods for binaural signals could be applied when wideband speech transmission is not available.

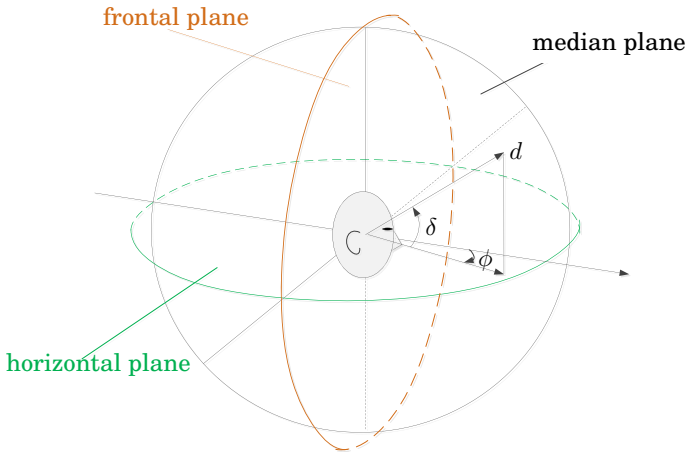

Figure 2.11. Coordinate system used to describe the position of a sound source with the listener placed at the origin. The position of a sound source can be defined by the azimuth, ϕ, the elevation, δ, and the distance, d.

Sound source localization

The position of a sound source is often described with the coordinate system shown in figure 2.11 [19]. When the listener is placed at the origin of the coordinate system, the position of a sound source can be defined by the three attributes the azimuth, ϕ, the elevation, δ, and the distance, d. The median plane cuts the head of the listener in two symmetrical halves, the horizontal plane is defined by the interaural axis and the lower margins of the eye sockets, and the frontal plane is orthogonal to the horizontal plane intersecting the interaural axis.

The existence of two ears is the main reason behind the ability of human listeners to identify the direction of a sound source. The auditory system analyses many temporal and spectral cues from the signals received by the ears. If a sound source is not located in the median plane, the sound signal arrives earlier to the nearer than to the farther ear. In other words, there is a time difference between the signals, which is usually referred to as the interaural time difference (ITD). In addition, the sound shadow of the head attenuates the signal on its way to the farther ear. This intensity difference between the two ears is called the interaural level difference (ILD). The maximum ITD, of about 700 μs, and the maximum ILD, of about 6 dB, are achieved when the sound source is located in the horizontal plane in the direction of $\phi = \pm 90°$.

ITD and ILD are considered the main cues for the localization of sound

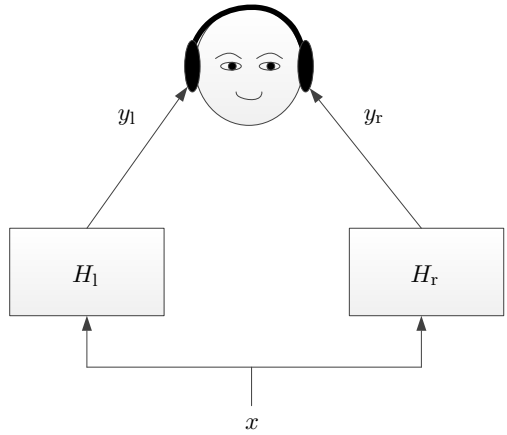

Figure 2.12. Binaural sound production through headphones. The left and the right channels are obtained by filtering the mono signal, x, with the HRTF filters H_l and H_r.

sources, but also the asymmetry of the outer ear, head and torso contribute to the directional hearing. For instance, the front-back separation is possible mainly because the pinna changes the spectral characteristics of the sound differently depending on whether the sound source is in front of or behind the head. Furthermore, the detection of the elevation angle is possible due to the asymmetry of the outer ear and the reflections from the shoulders.

The precision of the localization of a sound source, i.e. the localization blur, depends not only on the azimuth and elevation angles but also on the frequency and bandwidth of the sound signal. To generalize, the localization precision is better for low frequencies, whereas at around 1500 Hz, the precision is much worse because the signal wavelength is comparable to the size of the head, and the ITD is not a valid cue anymore. In addition, the localization blur is smallest when the sound source is in front of the listener, at position $\phi = 0°$. When the sound source is located at positions $\phi = \pm 90°$, the localization blur is at its maximum [20].

A head-related transfer functions (HRTF) can be used to describe how an ear receives a sound. The HRTF is a transfer function from a point sound source to the ear measured in a free field. Typically, the HRTFs are measured in unechoic chambers using a dummy head.

3D sound

3D sound refers to an attempt to reproduce sound through loudspeakers or stereo headphones to a listener creating an illusion of a natural envi-

ronment and sound sources.

With headphones, the perception of a sound depends directly on the signals that are brought to the ears. Without any differences or with only time and level differences, the sound is localized inside the head of the listener. This effect, called *lateralization* [21], is due to the fact that all the cues induced by the outer ears and the head are missing. Therefore, to create 3D sound with headphones, these cues should obviously be included in the signals. This can be achieved by processing the signals with the corresponding HRTFs, as shown in figure 2.12. The left channel, y_{l}, is obtained by filtering a mono input signal, x, through the left HRTF filter, H_{l}, and the right channel, y_{r}, is obtained by filtering the input signal, x, through the right HRTF filter, H_{r}, respectively [20].

Achieving the same spatial impression in loudspeaker listening, requires further processing. The HRTFs have to be modified to compensate the crosstalk from the other loudspeaker to the farther ear.

3. Digital speech transmission

The telephone was invented in 1876 by Bill Graham Bell. According to Oliver Lodge, Graham Bell's telephone

> "took the sounds emanating from the human voice, and sought to reproduce them at a distance by electrical means; so that at any, even great, distances, aerial vibrations could be reproduced and interpreted by the human ear, after the same fashion as the original aerial vibrations could be interpreted by an ear close to" [22].

Since the early days of the telephone, there have been limitations that have prevented this aim from being completely achieved. One of the limitations has been the bandwidth of the transmitted speech signal resulting in much worse speech quality than in face-to-face conversation. The human voice contains frequencies from 20 Hz to 20000 Hz, whereas the traditional telephone bandwidth, also called *narrowband*, only covers the range from 300 Hz to 3400 Hz. The limited bandwidth reduces speech quality [23] and intelligibility [24]. Furthermore, speaker recognition becomes more difficult due to the limited frequency band.

The used telephone bandwidth originates from analogue telephony, where the bandwidth was limited due to the characteristics of the transducers and other hardware, and due to the analogue frequency-division multiplexing with a frequency grid of 4 kHz [25]. Digital speech transmission was built around the same principles as the analogue system for compatibility reasons, and for decades, consumers were offered only narrowband speech transmission. Recently, also *wideband* speech transmission (50-7000 Hz) has become available in VoIP applications and gradually also in cellular networks. So far, HD voice has been launched on 41 mobile networks in 33 countries [26]. The transition from narrow-

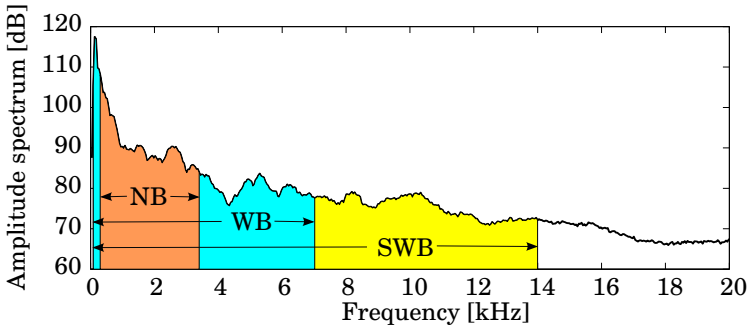

Figure 3.1. Average speech spectrum calculated from a 10-s long speech sample of a male speaker. The narrowband (NB) bandwidth (300-3400 Hz) is coloured red, the wideband (WB) bandwidth (50-7000 Hz) is coloured blue, and the superwideband (SWB) (50-14000 Hz) yellow.

band communications to wideband in cellular networks continues, but it will take years, because in addition to the networks, also the terminals have to be upgraded to wideband compatible ones. In addition to wideband speech transmission, also *superwideband* (50–14000 Hz) and *fullband* (20–20000 Hz) speech transmission are being developed [27]. An average speech spectrum calculated from a speech sample of a male speaker and the different telephone bandwidths are illustrated in figure 3.1.

3.1 Speech coding

At present, speech is mostly transmitted in digital format in telecommunication networks, such as the public switched telephone network (PSTN), digital cellular networks, and VoIP networks. For digital transmission, the analogue speech signal has to be represented in a digital form, and for this reason speech coding is needed. In other words, speech coding aims to represent the speech signal in a digital form with as few bits as possible while maintaining adequate speech intelligibility and quality for the application in mind. In addition to the bit rate and quality, speech coders can be characterized by their complexity and the delay they introduce [27].

The desired **bit rate** of a speech codec is determined by the channel capacity of the application. There is a trade off between the bit rate and the voice quality and intelligibility [28]. In some applications, having a coder with multiple bit rates is also desirable. The coder may change the bit rate on the fly depending on the available channel capacity [27].

The **quality** of a speech coder is usually expressed as a mean opinion score (MOS) value. It is a five-point scale from 1 to 5 (bad, poor, fair, good,

excellent) that is obtained from a subject as a result of a subjective listening test or estimated using an objective measure. It should be noted that MOS values obtained from separate listening tests or objective evaluations should not be directly compared with each other. Especially, care should be taken when comparing the quality of speech signals of different bandwidths.

The **complexity** of a speech coder is usually represented as the computational requirement (millions of instructions per second, MIPS) and memory consumption. The target is to minimize the complexity, as it directly affects the cost and energy usage of an application [27].

According to [29], in a telephone conversation, the end-to-end **delay** from the far-end user's mouth to the near-end user's ear is desired to be less than 150 ms. This is regarded as the limit for transparent interactivity, and greater delays hinder the conversation. However, a study reported in [30] suggests that a much bigger delay could be tolerated in a two-way conversation. The speech coder is one of the processing steps in the whole end-to-end processing chain, and it contributes directly to the delay.

3.2 Pulse code modulation

Pulse code modulation (PCM) coding became a coding standard for PSTN networks in the 1970's. PCM is a waveform coding method that aims to represent the time-domain speech waveform as accurately as possible. It uses a sampling rate of 8 kHz to sample analogue speech signals. The PCM speech bandwidth (300–3400 Hz) was adapted from analogue telephone systems, and it is specified in [1]. A non-linear quantization called A law (Europe, Africa, Australia, South America) or μ law (North America and Japan) is used. Both quantization laws are logarithmic such that more quantization levels are reserved for low amplitude values. The quantized amplitude values are then encoded with 8 bits, which results in the logarithmic PCM (log-PCM) coding with the 64-kbit/s bit rate, as specified in the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T) standard G.711 [31]. The log-PCM achieves a MOS value of 4.3 that is called the toll quality. Another variant of PCM is the adaptive differential PCM (ADPCM) at 16/24/32/40 kbit/s [32].

3.3 Narrowband speech in cellular networks

The first generation cellular networks, such as the Nordic Mobile Telephone (NMT) in Scandinavia and the Advanced Mobile Phone System (AMPS) in America, were analogue systems. Various standards existed around the world, and different mobile phones were needed for each standard. The speech quality in analogue systems is dependent on the signal quality. For example, the distance between the mobile phone and the base station impacts the quality directly so that when this distance increases the quality drops.

In the mid 1980's, the work on the European standard for a digital mobile cellular networks was started by the European Conference of Postal and Telecommunications Administrations (CEPT), and the work was later moved to the European Telecommunications Standards Institute (ETSI). The task was known as the Groupe Spécial Mobile (GSM), but the work was later renamed as Global System for Mobile Communications [33]. Since the commercial launch of GSM in the 1990's, the number of GSM and other mobile subscribers began to increase rapidly, reaching 5.6 billion mobile connections in 2011 [34].

Starting from the 1980's, the research focus in speech coding was on developing narrowband low-rate coders for cellular and military communication [35]. The starting point for the second generation coders was the narrowband PCM coded signal. The work on speech coding aimed at developing speech coding algorithms with a much lower bit rate than 64 kbit/s yet having adequate speech quality and tolerable complexity for mobile communications.

The first codec for the GSM, RPE-LTP (regular pulse excitation with long term prediction), is based on LPC analysis and operates at a bit rate of 13 kbit/s. The MOS value of the RPE-LTP codec is approximately 3.5, which is significantly lower than that of PCM. In addition to the RPE-LTP codec, many other speech codecs were standardized for GSM, Universal Mobile Telecommunications System (UMTS), and other 2G and 3G cellular networks. For example, the GSM Enhanced Full Rate (EFR) codec operating at 12.2 kbit/s, is based on an LPC-based method called algebraic code excited linear prediction (ACELP). Another example is a half rate coder called the vector sum excited linear prediction (VSELP) coder with bit rate of 5.6 kbit/s. Some of the well-known narrowband speech codecs are listed in table 3.1.

The adaptive multi-rate (AMR) coder standardized by ETSI is an example of speech coders having multiple bit rates [5]. It is based on ACELP technology and operates at eight different bit rates from 4.75 kbit/s to 12.2 kbit/s. The highest bit rate corresponds to the GSM EFR speech coder. The bit rate of the AMR coder adapts to the radio channel conditions so that when more channel capacity is available more bits can be used for speech coding, and consequently better speech quality is achieved.

Table 3.1. List of the most relevant narrowband (300-3400 Hz) speech codecs, their bit rates, and trend-setting MOS values [36].

Algorithm	Standard	Bit rate (kbit/s)	Quality (MOS)
log-PCM	G.711	64	4.3
ADPCM	G.726	16/24/32/40	toll
RPE-LTP	GSM FR	13	3.7
VSELP	IS-54	5.6	3.6
LD-CELP	G.728	16	4
CS-ACELP	G.729	8	4
ACELP	GSM EFR	12.2	4

3.4 Wideband and beyond

The first speech codec standardized for wideband speech for mobile communications was the Adaptive Multirate Wideband (AMR-WB) codec [37]. The AMR-WB was released in 2001 by the 3rd Generation Partnership Project (3GPP), and the same speech coder was also selected as a ITU-T recommendation, G.722.2. The AMR-WB codec is based on ACELP technology and operates at nine bit rates from 6.6 kbit/s to 23.85 kbit/s. The main difference to the GSM AMR codec is that it operates at a sampling rate of 16 kHz, which is required for the nominal range of 50–7000 Hz.

In mobile communications, the speech quality of PSTN was achieved with GSM EFR and AMR codecs. However, it was not until the AMR-WB codec that the narrowband PCM quality was finally exceeded. In [37], the perceived speech quality of the AMR-WB with bit rates starting from 8.85 kbit/s are superior to the quality of narrowband AMR at 12.2 kbit/s.

For wideband telephony in GSM or UMTS networks, either tandem-free operation (TFO) or transcoder-free operation (TrFO) is required. In TFO, the compressed wideband speech parameters are transmitted within the PCM 64 kbit/s bit stream. Wideband speech quality is achieved, but a 64 kbit/s bit rate is required. Another option is to use TrFO, where the

whole end-to-end link supports the same codec type, and transcoding is not needed. With TrFO, wideband speech is obtained with lower transmission rates.

The integration of the AMR-WB in 2G and 3G mobile networks is ongoing. The first operator to launch the AMR-WB for consumers in a 3G network was Orange in Moldova in September 2009 [7]. For consumers, the better speech quality achieved by AMR-WB coding is marketed as HD voice, and the number of operators worldwide offering HD voice in 3G networks is gradually increasing. However, the transition to wideband is time-consuming, since all terminals and networks have to be upgraded to wideband before full coverage is achieved. During the transition phase, the speech bandwidth of each phone call is dictated by the weakest link of the connection, i.e. the two terminals and the network between them. Furthermore, the speech quality may vary within a phone call due to both narrowband-to-wideband and wideband-to-narrowband handovers. In these situations, ABE methods can be utilized to narrow the quality gap between narrowband and wideband speech.

The next generation of mobile networks (4G), called Long Term Evolution (LTE), differs from the PSTN and 2G/3G networks by being a packet switched network instead of a circuit switched one. LTE was designed especially to offer higher data rates with lower latency for the increasing number of mobile data services. From the traffic point of view, even though mobile data volumes exceed voice volumes, conversational telephony will remain an important application in future mobile networks as well. Various solutions for voice over LTE (VoLTE) have been designed. For example, in GSM/WCDMA, the expected deployment strategy of VoLTE is three phased [38]. In the first phase, voice is transmitted in the legacy 2G/3G network, while only data is carried on LTE. This solution is called the circuit-switched fallback. The next phase uses IP multimedia system (IMS) based on VoIP solutions that enables handover from IMS-based VoIP to circuit switched speech when the user equipment is running out of VoIP coverage, and vice versa. Finally, in the third phase of the deployment strategy, all calls are made over packet-switched networks.

When the LTE was introduced by the 3GPP, the speech codecs were inherited from the UMTS [39]. The suitability of the default codecs, AMR and AMR-WB, have been tested for packet-switched conversational multimedia applications in [40]. However, VoLTE offers new possibilities for even wider speech bandwidth, and consequently for better speech quality.

4. Speech quality and intelligibility

In a telephone conversation, the acoustic signal is perceived by the near-end user's ear, causing an auditory event in the brain, which results in a description of the sound [41]. Both the prior knowledge of the communication system and the emotions of the listener affect the quality judgement [42]. In addition, both the content and the form of the signal are analysed by the listener, and in telecommunications, speech quality usually refers to the form of the speech signal, i.e. the acoustic signal, although both content and individual factors affect the quality perception [43]. Speech quality is complex to define. According to [44], speech quality encompasses attributes such as naturalness, clarity, pleasantness, and brightness. Quality relates to "how" a speaker produces an utterance. Intelligibility, on the other hand, is "what" is been said, and it is not a dimension of speech quality. A similar definition of speech quality is also used in this thesis. There are, however, other ways to define speech quality, where intelligibility is considered as a dimension of speech quality [45]. In speech transmission, the minimum requirement is that speech is intelligible enough, that what is said is understood. As the intelligibility increases, other factors like naturalness and recognizability of the voice become more important.

During the last decades, many new speech coding and transmission systems have been introduced. Besides the traditional narrowband (300–3400 Hz) speech coding and transmission, also wideband (50–7000 Hz) and superwideband (50–14000 Hz) speech transmission have been deployed in many telephone applications. In addition, together with VoIP applications, the number of different degradation types in speech signals, such as packet loss, has increased. As a result, the need for new effective and reliable methods for the evaluation of speech transmission systems, coding standards, and speech enhancement methods has increased.

Speech quality can be assessed by subjective or objective methods. In

subjective (auditory, perceptual) methods, the evaluation is made by asking people's opinions on speech sounds and applying statistical analysis to the data. Objective (instrumental) methods, on the other hand, are computational measures that try to predict the subjective speech quality.

4.1 Subjective quality evaluation

Subjective sound quality assessment can be categorized into four groups as shown in table 4.1 [46, 47]. The vertical categorization is made between *utilitarian* and *analytic* methods. In utilitarian methods, the subjects evaluate the integral quality using a one-dimensional scale. These tests can be used to compare varying conditions resulting from different speech coding algorithms. In analytical test methods, the subjects evaluate certain features of the perceived sound. The subjects may be asked to assess one certain feature using a one-dimensional scale or a number of quality features with several scales. The horizontal classification of the quality test methods in table 4.1 is made between *subject-oriented* methods and *object-oriented* methods. The former focus on gathering information on human perception, whereas the latter are used to evaluate the quality of a certain system.

Table 4.1. Subjective listening quality tests following [46].

	Subject-oriented	Object-oriented
Utilitarian	Psychoacoustic research	Sound quality assessment
Analytical	Audiological evaluation	Diagnostic listening tests

The evaluation methods of speech transmission systems are mainly object-oriented. The overall speech quality is often analysed with utilitarian methods. For example, in audio standards, such as ITU-T P.800, P.830 and P.805, most of the methods are utilitarian and univariate tests. However, also analytical methods may be used, and they could result in more in-depth understanding of speech quality in digital transmission systems. An example of such tests is the ITU-T Recommendation P.835, that is a standard for evaluating speech transmission systems that include a noise suppression algorithm [48]. In the test, the subjects are asked to assess the speech signal, the background noise, and the overall effect separately.

Subjective evaluation of speech quality is needed when reliable objective measures do not exist. Also, subjective evaluation may be employed when evaluating the overall quality or complex parameters of speech quality.

However, subjective evaluation is not very effective, because the collection of each data point requires that a subject grades the performance of a sample. The resulting data is also prone to variance, since subject's personal opinion is involved. To improve the reliability of subjective evaluation, rigorous testing procedures are required. Formal subjective tests following formal methods and standards are more time-consuming to arrange than informal tests. However, they provide more reliable and repeatable results. In addition, formal statistical analysis gives information on the quality of data.

The selection of the listeners for a subjective test depends on the test type in use. Naive (untrained) test subjects are usually involved in utilitarian tests, where the objective is to find out the opinion of the average telephone user. According to [49], a naive subject has not been involved in work connected with assessment of telephone circuits, has neither participated in any subjective test for at least the previous six months nor in any listening opinion test for at least one year, and has not heard the same sentence lists before. The subjects should not have any kind of hearing impairment, and their mother tongue should be the same as the language in the test [47]. An alternative for naive test subjects is to use expert (trained) listeners. For instance, in analytical methods, the results of the test are more reliable if the subjects have been trained for the task.

4.1.1 Listening-only tests

The speech quality of a communication system is most often assessed by listening-only tests. Usually this refers to using a pre-recorded and processed set of speech samples that is presented to the subjects, for example, through headphones, and the subjects are asked to assess the quality using a predefined scale given by the experimenter. An advantage of these methods is that the evaluation is subjective, i.e. subjects listen to real speech samples and grade them according to their personal opinion. A drawback is that listening tests are rather directed, meaning that test design factors and rating procedures affect the subjects perception process. For example, the samples are pre-recorded and usually quite short (one sentence, for example), and thus a single artefact in a sample may dictate the entire grading of the sample. Furthermore, the subject focuses mainly on the form of the acoustic signal and not the content because he or she is placed only in a listening context, which is not natural in a normal telephone conversation situation.

ITU-T recommendation P.800 describes several test methods and category rating scales for listening opinion tests that are often used to evaluate speech transmission systems [49]. The recommendation also describes reference conditions that are important if two tests arranged in different laboratories or at different times are to be compared with each other.

Absolute category rating (ACR) is perhaps the most widely used test to assess the overall speech quality. In the test, the subjects are asked to assess the speech quality using the 5-point mean opinion scale (MOS) shown in table 4.2. For instance, the performance of speech codecs is usually evaluated with ACR tests. Listening effort and loudness preference may also be evaluated using a 5-point scale from 5 to 1 [49].

Table 4.2. Mean opinion score according to ITU-T P.800 [49].

Score	Quality of the speech
5	excellent
4	good
3	fair
2	poor
1	bad

The degradation category rating (DCR) test is designed to compare small degradations in speech quality compared to a reference system. The subject is asked to evaluate a degraded signal compared to a reference using the degradation mean opinion score (DMOS) shown in table 4.3. The reference signal is always presented first to the subjects followed by the test signal.

Table 4.3. Degradation mean opinion score according to ITU-T P.800 [49].

Score	Degradation is
5	inaudible
4	audible but not annoying
3	slightly annoying
2	annoying
1	very annoying

The comparison category rating (CCR) test is another test where two signals are compared with each other. In the CCR test, the latter signal is rated compared to the former with the comparison mean opinion score (CMOS) shown in table 4.4. This test can also be applicable to assess small differences between two or more systems. The results of a CCR test provide information on which sample is better and by how much.

Table 4.4. Comparison mean opinion score according to ITU-T P.800 [49].

Score	Quality of the second compared to that of the first
3	much better
2	better
1	slightly better
0	about the same
-1	slightly worse
-2	worse
-3	much worse

Binary paired comparison can be employed in speech quality assessment as well. The test signals are presented to the listeners in pairs, and the subjects are asked to choose the one they prefer. The paired comparison test is suitable when the differences between the conditions are small.

4.1.2 Conversational tests

Compared to listening-only tests, conversational tests are a step closer to a real conversation situation. Everyday telephone communication is mostly conversational, and the quality perception is undirected and individual. The listener pays attention to features that the individual person considers relevant for the communication situation. For example, the telephone user may consider important features like intelligibility, loudness, listening effort, or naturalness. The significance of semantic information increases as the user is placed in a conversational context where listening, talking, double-talk, and periods of mutual silence alternate. Conversational tests are, however, more expensive and time-consuming to arrange and therefore quite rare.

A newish ITU-T recommendation P.805 describes a subjective test to assess conversational quality over a telephone transmission [50]. The test can be used to evaluate either a specific source of degradation, such as delay or echo, or the overall quality of a transmission system. The target is to have as realistic a communication environment as possible, where two people are having a true spontaneous conversation over a telephone system. In practice, two subjects at a time are placed in separate sound proof rooms. The subjects can be experts, experienced, or naive participants, depending on the purpose of the test. During the test, the subjects have a conversation on a given topic, or task, and then give their opinion on the voice quality. There can be simulated noise environments either at

one end of the conversation or at both ends.

4.1.3 Field tests

The speech quality experienced by the end user is affected by the entire speech link, from the far-end to the near-end user. Field tests offer the most realistic environment for assessing speech quality of transmission systems. For example, the speech quality of a mobile phone could be assessed during true phone calls. In practice, field tests are rare and expensive to arrange, especially during the development process of a system, and therefore quality tests are usually arranged in laboratory conditions. Laboratory conditions are also easier to design such that the test is repeatable.

4.2 Objective quality evaluation

Objective (instrumental) quality assessment methods are computational tools that have been developed to evaluate the quality of speech signals and communication systems. Some of the measures analyse only certain aspect of the signals, whereas others try to model human perception more precisely. They are fast, repeatable, and automatic tools compared to time-consuming subjective testing, but still they are only models and can not replace subjective testing completely.

Objective quality assessment methods are often classified as parameter-based and signal-based models [43, 47]. Parameter-based models, such as the E model [51], rate the integral quality of an entire transmission path. They provide information on the whole network and are useful, for example, in network planning. Signal-based methods are more useful in the field of speech enhancement. In such methods, the quality score of a system is computed directly from the degraded signal and the original reference signal, or alternatively only from the degraded signal. Some of the methods measure only one feature of the sound, whereas others model the overall quality through a perceptual model.

The simplest objective measures include time-domain and spectral-domain measures, such as SNR ratio, mean square error, or spectral distance measures. These measures can be useful in many cases, but are not very good predictors of subjective speech quality. Therefore, perceptual measures have also been designed.

The perceptual evaluation of speech quality (PESQ), standardized in ITU-T Recommendation P.862 [52], can be used to evaluate narrowband speech codecs or end-to-end quality. Both the degraded and a reference signal are necessary. The model calculates several delays between the two signals and compares them using a perceptual model. The result is an objective quality score. This score can be mapped to a listening quality MOS that allows linear comparison with the MOS. An extension of the PESQ for wideband signals is presented in the recommendation P.862.2 [53]. An upgrade for the PESQ is called the perceptual objective listening quality analysis (POLQA) and is presented in the recommendation ITU-T P.863 [54]. The POLQA supports narrowband, wideband and superwideband speech signals, and it is intended to cover most of the telephone network scenarios.

4.3 Intelligibility tests

The first subjective tests focused on measuring intelligibility through subjective tests. In articulation or word tests, intelligibility is measured as the percentage of correctly recognized speech sounds at the receiving end of a system. Either short segments of speech, such as monosyllables, or complete words can be used. In rhyme tests, introduced by Fairbanks [55], rhyming words are used instead. The diagnostic rhyme test (DRT) consists of 96 rhyming word pairs that differ in their initial consonant [56]. The subject hears one word at a time and identifies the word he or she thinks he or she heard from the pair of words listed. An error rate is calculated from the results. The modified rhyme test (MRT) contains 50 word lists of six one-syllable words, differing in either the initial or the final consonant [57]. The subject hears one word at a time and chooses from the list the word he or she thinks he or she heard. The result of the test is an error rate of correctly identified words.

Speech intelligibility tests with speech segments longer than single words take better into account impairment of continuous speech. Speech reception threshold (SRT) is a test to find a presentation level for test speech necessary for a listener to understand the speech correctly a specified percentage of the time, usually 50%. In the SRT, complete test sentences are presented either in silence or in the presence of a reference noise signal.

5. Artificial bandwidth extension of speech

5.1 Background

ABE methods for speech attempt to regenerate the frequency content that is lost due to narrowband speech coding. Usually narrowband speech bandwidth refers to the frequency range from 300 Hz to 3400 Hz that is used in many existing speech coding standards, e.g. in PCM [31], the AMR codec [5], or the G.729 codec [58]. The ABE method typically doubles the 8-kHz sampling rate of narrowband speech to 16 kHz and adds new frequency content to the signal. Bandwidth extension towards high frequencies creates new content in the frequency band from 3.4 kHz (or 4 kHz) to 7 kHz (or 8 kHz). There are also bandwidth extension methods towards low frequencies, 50–300 Hz [59, 60, 61]. Although the naturalness of voice is degraded due to the missing low frequency content, these methods are not studied in this thesis. High frequencies are more important from the point of view of speech intelligibility. Furthermore, the reproduction of the low frequencies with small earpieces of mobile devices is not always possible.

Bandwidth extension methods can be further classified as artificial methods and methods with side information. The word "artificial" refers to algorithms that attempt to regenerate the lost frequency content utilizing *only* the information available in the narrowband signal, i.e., no information about the missing frequencies is transmitted. Since the extension is solely based on the narrowband signal, these methods can be implemented at the receiving end of the transmission channel. There are also methods that are not artificial but utilize transmitted side information in the extension procedure [9, 10, 11]. These methods hide some side information related to the missing frequency band in the narrowband sig-

nal. The methods are independent and can be used with any narrowband speech codecs. A drawback of the bandwidth extension methods with side information is that their exploitation requires that the same method is supported at both ends of the communication link. On the other hand, the performance of such methods can obviously be superior to that of the artificial bandwidth extension methods, as was stated in [9, 11]. Bandwidth extension with side information can be also implemented as a part of speech coding, as in the AMR-WB codec [6] or in the G.729.1 codec [62]. The motivation for utilizing bandwidth extension techniques in speech codecs is to obtain better speech quality at very low bit rates, rather than to overcome limitations due to narrowband speech coding or to enhance transmitted narrowband speech.

The focus of this section, and of the whole thesis, is on ABE methods that aim to regenerate the signal at high frequencies. The missing frequency band from 4 to 8 kHz, i.e. the extension band, is therefore referred to as the *highband*.

5.1.1 Correlation between narrowband signal and the missing highband

Typically, ABE methods are data-driven algorithms that utilize true wideband references in the training of the extension procedure. They are built on the assumption that the narrowband signal and the missing highband are correlated and the narrowband signal contains enough cues to regenerate the missing highband. Especially for voiced speech, the correlation originates from the low-pass characteristics of the excitation signal.

The dependency between the narrowband signal and the missing highband has been addressed in [63, 64, 65, 66]. The upper bound on the achievable quality of a memoryless bandwidth extension system was discussed in [64]. In their study, the mutual information between the features of the narrowband speech signal and the representation of the missing frequency band was evaluated with respect to an objective distance measure, a mean log spectral distortion (LSD), between the ABE output and the true wideband signal. The results indicate that to minimize the LSD, the narrowband features should be selected so that the mutual information (MI) is maximized.

The ratio between the MI of the narrowband and the highband, and the entropy of the highband was used in [63] to measure the uncertainty of the highband envelope given the narrowband. The dependency was found to

be relatively small, and the authors conclude that a bandwidth extension method with a memoryless mapping may perform well, not because of an accurate prediction of the highband, but because the signal bandwidth is extended such that the signal sounds pleasant. As an interpretation for the low MI between the narrowband and highband, it was explained in [65] that instead of one-to-one mapping, the narrowband and highband spectral envelopes have a one-to-many relationship.

The dependency analysis between the narrowband and the highband was further extended in [66, 67] by investigating the role of speech memory in increasing the certainty of the highband. The results showed that the certainty is increased as measured by the ratio of MI to the highband entropy, and the bandwidth extension methods can benefit from a short-term memory.

5.2 General model for artificial bandwidth extension

A general model for ABE that most of the ABE algorithms follow is shown in figure 5.1. The input signal, s_{nb}, is a narrowband signal with a sampling rate of 8 kHz. Through interpolation and lowpass filtering the signal is up-sampled to 16 kHz. The resulting signal, s_{lo}, is a wideband signal with narrowband content. A feature set is extracted from the input signal and used to estimate parameters for highband shaping. The excitation for the highband is first created to regenerate the highband signal. The highband signal, s_{hi}, is obtained after shaping the excitation utilizing the estimated shaping parameters. Finally, the ABE output, s_{abe}, is obtained by adding the generated highband signal and the up-sampled and lowpass filtered original narrowband signal.

The general model for ABE methods presented in figure 5.1 can be further modified to describe methods that are even more specifically based on the LPC-based speech production model. A majority of the ABE methods in the literature more or less follows the algorithm structure shown in figure 5.2. The highband excitation signal is obtained by extending the narrowband LPC residual signal to the highband. Additionally, the LPC coefficients for the highband envelope are also extended, and they are used in the LPC synthesis to produce the highband signal.

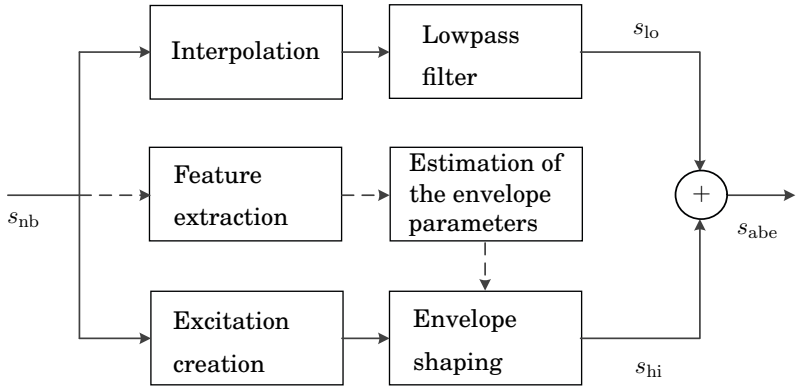

Figure 5.1. General model for ABE. The input signal, s_{nb}, is a narrowband signal with a sampling rate of 8 kHz. The ABE output, s_{abe}, with a doubled sampling rate of 16 kHz, is obtained by adding the up-sampled narrowband signal, s_{lo}, and the regenerated highband signal, s_{hi}.

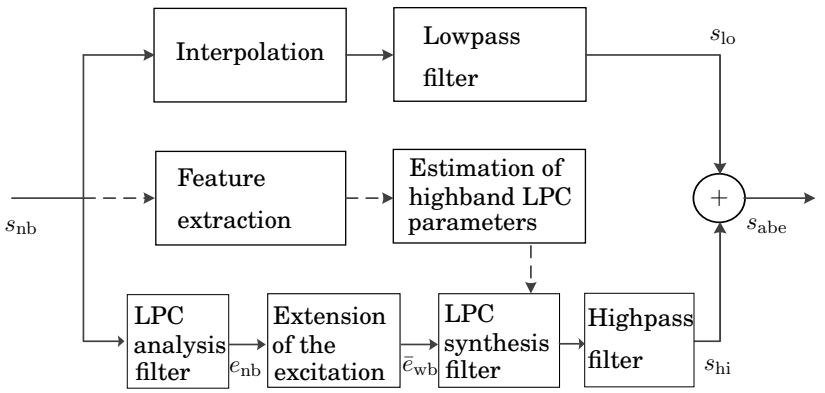

Figure 5.2. ABE based on LPC. The input signal, s_{nb}, is a narrowband signal with a sampling rate of 8 kHz. The ABE output, s_{abe}, with a doubled sampling rate of 16 kHz, is obtained by adding the up-sampled narrowband signal, s_{lo}, and the regenerated highband signal, s_{hi}.

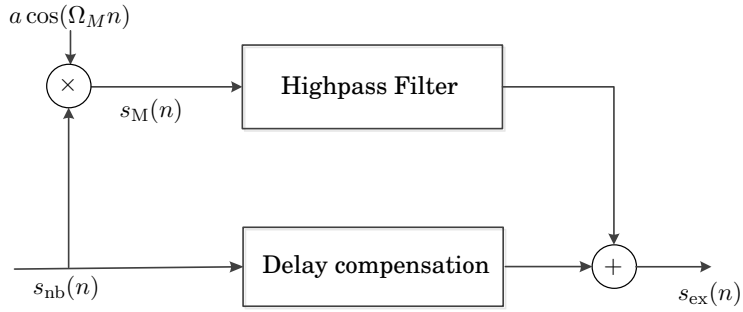

Figure 5.3. Extension of the excitation by cosine modulation. A delayed version of the original narrowband signal, $s_{\mathrm{nb}}(n)$ is added to the modulation output, $s_{\mathrm{M}}(n)$, that has been highpass filtered. As a result, a wideband excitation signal, $s_{\mathrm{ex}}(n)$ is obtained. Ω_{M} is the modulation frequency, and a is selected from $a\epsilon\{1,2\}$ so that the power of the excitation signal is correct.

5.3 Extension of the excitation signal

The techniques to create the spectral fine structure for the highband are generally called excitation extension methods. The word *excitation signal* originates from the source-filter model of the speech production mechanism, and in many ABE algorithms the extension of the excitation refers directly to the LPC residual signal, as shown in figure 5.2. However, the extension of the excitation may cover also spectral widening techniques used by such ABE methods that widen and shape the spectrum without specifically exploiting the source-filter model, as shown in figure 5.1. In addition, the highband excitation can be derived directly from the narrowband signal [68, 69] or more popularly from the narrowband LPC residual [70, 71, 72].

Non-linear processing applied to the narrowband excitation signal, $s_{\mathrm{nb}}(n)$, is used in many ABE algorithms as a technique to extend the excitation signal [68, 73, 74, 75, 76, 77]. The most popular non-linear functions include a quadratic function, $(s_{\mathrm{nb}}(n))^2$, a cubic function, $(s_{\mathrm{nb}}(n))^3$, and a fullwave rectifier, $|s_{\mathrm{nb}}(n)|$. These non-linear functions produce harmonic distortion components without a pitch estimation. A disadvantage of non-linear functions is that they produce highband excitation having a varying amplitude spectrum. Therefore, some sort of spectral flattening is needed.

Excitation extension can also be implemented through time-domain modulation of the narrowband excitation signal, which corresponds to spectral translation and folding methods [78]. Modulation techniques produce shifted copies of the original subband spectrum in the missing frequency

range. A block diagram of the modulation with a real-valued cosine function is shown in figure 5.3. A delayed version of the original narrowband signal, s_{nb} is added to the modulation output, $s_{\mathrm{M}}(n)$, that has been highpass filtered. The modulation function is of the form

$$s_{\mathrm{M}}(n) = s_{\mathrm{nb}}(n) \cdot a\cos(\Omega_{\mathrm{M}} n), \tag{5.1}$$

where Ω_{M} is the modulation frequency, and a is selected from $a\epsilon\{1, 2\}$ so that the power of the excitation signal is correct. In spectral translation, the modulation frequency is fixed, and it produces a shifted copy of the original spectrum in the higher frequency range [79, 78]. If the frequency band that is to be copied is from Ω_{lo} to Ω_{up}, the modulation frequency, Ω_{M}, is given as

$$\Omega_{\mathrm{M}} = \Omega_{\mathrm{up}} - \Omega_{\mathrm{lo}}. \tag{5.2}$$

The cut-off frequency of the highpass filter is Ω_{up}. As a result, the output signal, $s_{\mathrm{wb}}(n)$, is a sum of the original signal and its shifted copy from Ω_{up} to $\Omega_{\mathrm{up}} + \Omega_{\mathrm{M}}$. The extension starts right where the bandwidth of the original signal ends.

Spectral folding (or mirroring) is a special case of modulation and corresponds to modulation by the Nyqvist frequency, $\Omega_{\mathrm{M}} = \pi$, (i.e. 8 kHz in narrowband telephony). The output of spectral folding is a mirror image of the narrowband spectrum in the highband. It has the same effect as up-sampling the signal by two:

$$s_{\mathrm{wb}} = s_{\mathrm{nb}}(n) + s_{\mathrm{nb}}(n)(1 + (-1)^n). \tag{5.3}$$

Spectral folding can be applied to the narrowband signal (for example [68, 69]) or to the LPC residual signal (for example, in [70, 80, 81, 71, 82, 72, 83]). Spectral folding produces frequency components almost up to 8 kHz, but on the other hand, there is a spectral gap at around 4 kHz due to the original telephone bandwidth of 300–3400 Hz. The gap could be avoided by folding the spectrum already at 3.4 kHz. However, it has been reported that spectral gaps of moderate bandwidth have almost an inaudible effect in perception of speech [78]. A disadvantage of the modulation techniques with a fixed frequency is that the harmonic structure is not preserved in the highband. On the other hand, the human ear is not very sensitive to the harmonic structure at high frequencies as it is at low frequencies. In [15], the correction of highband harmonic structure was found to be unimportant for the perceived quality of ABE processed speech. Therefore, for example, spectral folding has been a popular

choice for excitation extension in ABE methods. The harmonic structure of the excitation signal can be preserved by utilizing an adaptive modulation frequency that is dependent on the current pitch frequency [79]. The pitch adaptive modulation has been implemented both in the time domain [84, 64] and in the frequency domain [85].

Sinusoidal synthesis has also been used to create a harmonic structure in the missing frequency band [86, 87, 88]. For sinusoidal synthesis, the harmonic structure for the highband is created by a bank of oscillators with frequencies, amplitudes, and phases that are determined from the narrowband speech. In sinusoidal synthesis, no spectral flattening is needed, since the spectral envelope can be created directly through sinusoidal amplitudes.

An excitation extension method of modulated noise is motivated by the fact that the harmonic structure becomes more noisy above 4 kHz. Furthermore, in the frequency band above 4 kHz the resolution of human hearing starts to worsen and the pitch periodicity is perceived through the time-domain envelope of the bandpass speech signal [89]. Therefore, the excitation can be extended by modulating highband noise with the time-domain envelope of the bandpass (2.5–3 kHz) signal [90, 91, 92].

5.4 Extension of the spectral envelope

The extension of the spectral envelope from the narrowband to the highband is usually made based on some model that maps a narrowband feature vector onto the wideband envelope. The feature vector may consist of features that describe the shape of the envelope directly or some other features that are related to the temporal waveform.

5.4.1 Features

A set of typical features used in ABE methods is given in the following.

Linear predictive coding (LPC) prediction coefficients can be utilized to describe the spectral envelope.

Line spectrum frequency (LSF) also known as line spectrum pair (LSP), is an alternative representation for LPC coefficients [93]. LSF decomposition is used in many speech coding standards as being an efficient quantization technique. The LSFs are defined as the roots of the polynomials

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)}A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)}A(z^{-1}), \end{aligned} \tag{5.4}$$

where $A(z)$ is the LPC inverse filter.

Cepstral coefficients, $c(n)$, are computed as an inverse discrete Fourier transform of the logarithm of the power spectrum of the signal, $s(n)$, [13]:

$$c(n) = \mathcal{F}^{-1}\{\log|\mathcal{F}\{s(n)\}|\} \tag{5.5}$$

Alternatively, cepstral coefficients that are transformed from the LPC coefficients are also often used in ABE methods. Cepstral coefficients $[c_0, c_1, \ldots, c_P]$ are calculated from the LPC coefficients $[a_0, a_1, \ldots, a_P]$ using a recursive equation:

$$\begin{aligned} c_0 &= \sigma^2 \\ c_i &= -a_i - \textstyle\sum_{n=1}^{i-1} \frac{n}{i} c_n a_{i-n}, \end{aligned} \tag{5.6}$$

where σ is the rms value of the signal that is normalized to 1.

Mel frequency cepstral coefficients (MFCC) are computed as cepstral coefficients, but instead of using a linear frequency scale, a perceptually motivated mel frequency scale is used [94].

Energy-based features include different versions of frame energy. Energy can be calculated from the entire narrowband signal or from subbands. Frame energy for a time-domain frame $s(n)$ with a frame length N is defined as

$$x_{\mathrm{e}} = \sum_{k=0}^{N-1} (s(n)^2), \tag{5.7}$$

Zero crossing is a traditional feature for voiced/unvoiced clustering. It is calculated from the time-domain frame $s(n)$:

$$x_{\mathrm{zc}} = \frac{1}{N-1} \sum_{k=1}^{N-1} \frac{1}{2} |\mathrm{sign}(s(k-1)) - \mathrm{sign}(s(k))|, \tag{5.8}$$

where N is the frame length, and the sign operation is defined as

$$\mathrm{sign}(x) = \begin{cases} +1, & \text{if} \quad x > 0 \\ 0, & \text{if} \quad x = 0 \\ -1, & \text{if} \quad x < 0. \end{cases} \tag{5.9}$$

Gradient index was originally introduced in [95]. It is a feature for describing voiced/unvoiced characteristics of a speech frame $s(n)$:

$$x_{\mathrm{gi}} = \frac{1}{10} \frac{\sum_{k=1}^{N-1} \Psi(k)|s(k) - s(k-1)|}{\sqrt{\sum_{k=0}^{N-1} (s(k))^2}}, \tag{5.10}$$

where N is the frame length, $\Psi(k) = 1/2|\psi(k) - \psi(k-1)|$, and ψ is the sign of the gradient $s(k) - s(k-1)$.

Spectral flatness is a frequency-domain feature calculated from the power spectrum $|S(e^{\mathrm{j}\Omega_i})|$:

$$x_{\mathrm{sf}} = \log_{10} \frac{\sqrt[N_i]{\prod_{i=0}^{N_i-1} |S(e^{\mathrm{j}\Omega_i})|^2}}{\frac{1}{N_i} \sum_{i=0}^{N_i-1} |S(e^{\mathrm{j}\Omega_i})|^2}, \tag{5.11}$$

where N_i is the length of the discrete Fourier transform (DFT), $e^{\mathrm{j}\Omega_i}$ is the ith DFT frequency, j being the imaginary unit, and $e = 2.7182$ the base of the natural logarithm [96].

Centroid of the power spectrum is a frequency-domain feature that results in higher values for unvoiced speech than for voiced speech. It is defined as:

$$x_{\mathrm{sc}} = \frac{\sum_{i=0}^{N_i/2} f(i)|S(e^{\mathrm{j}\Omega_i})|}{(N_i/2+1) \sum_{i=0}^{N_i/2} |S(e^{\mathrm{j}\Omega_i})|}, \tag{5.12}$$

where $|S(e^{\mathrm{j}\Omega_i})|$ is the power spectrum, $f(i)$ refers to the frequency in the ith DFT bin, and N_i is the length of the DFT [96, 72].

5.4.2 Distance measures

Distance measures play an important role in many ABE algorithms. Typically, the distance measure between two spectral envelopes is needed in the training phase of a Gaussian mixture model, codebook, or neural network. In codebook-based methods, a distance measure is also utilized in the selection of a codeword for each frame. Similar measures have also been used in objective quality evaluation of ABE methods. Most of the spectral distance measures are mean square error measures that are applied either to the spectrum directly or to a parametric representation, such as the LPC envelope. Examples of typical distance measures are:

Logarithmic spectral distortion (LSD) is mean square error in dB:

$$d_{\mathrm{LSD}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} (20 \log_{10} \frac{\sigma}{|A(e^{\mathrm{j}\Omega})|} - 20 \log_{10} \frac{\hat{\sigma}}{|\hat{A}(e^{\mathrm{j}\Omega})|})^2 \mathrm{d}\Omega, \tag{5.13}$$

where $\frac{1}{A(e^{j\Omega})}$ is the LPC envelope of the original highband speech, $\frac{1}{\hat{A}(e^{j\Omega})}$ is the LPC envelope of the estimated highband speech, and σ and $\hat{\sigma}$ are the respective relative gains.

Cepstral distance is a commonly used distance measure in speech processing that is calculated directly from p cepstral coefficients of the original wideband signal, $\mathbf{c}$, and the estimated wideband signal, $\hat{\mathbf{c}}$:

$$d_{\mathrm{CEPS}} = \sum_{i=1}^{p} (c_i - \hat{c}_i)^2. \tag{5.14}$$

5.4.3 Codebook mapping

Codebook mapping was one of the first approaches for highband envelope estimation in the ABE field [97, 70, 98, 80, 99, 81, 100, 91, 76]. Put simply, a narrowband codebook consists of a list of codewords, i.e. narrowband feature vectors, and corresponding highband envelopes. The feature vector computed from the input speech is compared with each of the codewords, and the best match is selected. The best match is decided on the bases of an error measure, for example the mean square error, between the input vector and the codebook entries.

The basic codebook mapping can be improved with modifications as presented in [99, 100, 91]. Separate codebooks were constructed for unvoiced and voiced fricatives in [100], which improved the performance of the algorithm measured in terms of spectral distortion (SD). Similar results were obtained in [99], where a split codebook for voiced and unvoiced speech sounds improved the performance. In addition, a codebook mapping with interpolation, i.e. where the highband envelope is calculated as a weighted sum of the most probable codebook entries, was reported to enhance the extension quality in [99, 91]. Furthermore, in [91], codebook mapping with memory was implemented by interpolating the current envelope estimate with the envelope estimate of the previous frame.

Typically, both the narrowband feature vectors and the highband codewords directly represent the spectral envelopes through LPC coefficients ([97, 70, 80]) or LSFs ([98, 100, 91]). MFCCs have also been used in [81]. In [76], a slightly different approach was chosen, where an estimate for highband energy is first calculated after which the energy is mapped onto the highband spectral envelope codebook.

5.4.4 Linear mapping

Linear mapping is utilized in [73, 99, 82] to estimate the spectral envelope of the highband. In linear mapping, the input narrowband envelope is characterized by a vector of parameters $x = [x_1, x_2, \ldots, x_n]$ and the wideband envelope to be estimated by another vector $y = [y_1, y_2 \ldots, y_m]$. For example, LPC prediction coefficients or LSFs can be used as input and output parameters. The linear mapping between the input and output parameters is then denoted as

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{5.15}$$

where the matrix $\mathbf{W}$ is obtained through an off-line training procedure with the least-squares approach that minimizes the model error $\mathbf{y} - \mathbf{W}\mathbf{x}$ using a training data with narrowband envelope parameters $\mathbf{X}$ and corresponding highband parameters $\mathbf{Y}$:

$$\mathbf{W} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}. \tag{5.16}$$

To better reflect the non-linear relationship between the narrowband and highband envelopes, modifications for the basic linear mapping technique have been presented. Instead of using a single mapping matrix, the mapping can be implemented by several matrices. In [82], speech frames are clustered into four clusters based on the first and the second reflection coefficients, and a separate mapping matrix is created for each cluster. The algorithm in [82] hence utilizes hard-decision clustering, whereas a soft decision scheme is implemented in [73], where the clustering is performed through vector quantization of the input vector, $\mathbf{x}$, and the final output vector, $\mathbf{y}$, is formed as a weighted sum of the mappings obtained for all clusters.

The evaluation of the performance of the ABE methods with linear mapping is rather concise. In [73], based on an informal listening test conducted in the Japanese language, the quality of narrowband speech was better than the extended speech. Objective analysis was also provided, showing that SD for piecewise-linear mapping was smaller than for codebook mapping and neural network approaches. On the other hand, the objective comparison given in [99] indicates better performance for codebook mapping compared to linear mapping.

5.4.5 Gaussian mixture model

In linear mapping, only linear dependencies between the narrowband spectral envelope and the highband envelope are exploited. Non-linear dependencies can be included in the statistical model by utilizing Gaussian mixture models (GMM). GMM is a parametric model for modelling high-dimensional probability density functions (PDF) of a random variable.

In ABE methods, the GMM is typically utilized to model the joint PDF between two random variables, $\mathbf{x}$ and $\mathbf{y}$. The GMM PDF for variables $\mathbf{x} = [x_0 ... x_{b-1}]$ and $\mathbf{y} = [y_0 ... y_{d-1}]$ is represented as a weighted sum of L Gaussian component densities f_{G}:

$$f_{\mathrm{GMM}}(\mathbf{x}, \mathbf{y}) = \sum_{l=i}^{L} \rho_l f_{\mathrm{G}}(x, y; \mu_{\mathbf{x},\mathbf{y},l}, \mathbf{V}_{\mathbf{x},\mathbf{y},l}), \tag{5.17}$$

where L is the number of individual Gaussian components, ρ_l is the lth mixture weight, $\mu_{\mathbf{x},\mathbf{y},l}$ is the mean vector and $\mathbf{V}_{\mathbf{z},l}$ is the covariance matrix. These GMM parameters can be estimated from training data through an iterative training procedure, such as the expectation maximization (EM) algorithm.

The GMM is utilized directly in envelope extension to estimate wideband LPC coefficients or LSFs from corresponding narrowband parameters [101, 90]. The performance of the GMM-based spectral envelope extension was then enhanced by using MFCCs instead of LPC coefficients [87, 102]. Furthermore, the GMM mapping with memory further results in better performance in terms of LSD and PESQ [103, 104]. In addition to using GMMs directly in envelope extension, they can be used in HMM-based ABE algorithms as well. These methods are discussed in the next section in more detail.

The advantage of the GMM in envelope extension methods is that they offer a continuous approximation from narrowband to wideband features compared to the discrete acoustic space resulting from vector quantization. Better results were reported for GMM-based methods compared to codebook mapping in [101] and [90] in terms of SD, cepstral distance, and a paired subjective comparison.

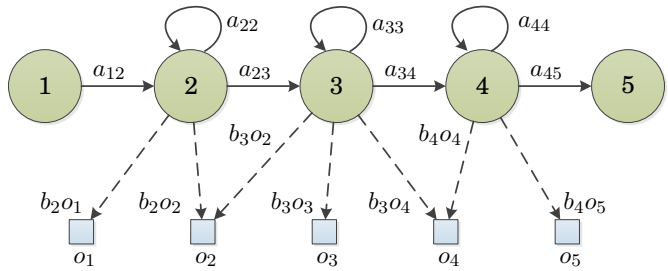

Figure 5.4. Hidden markov model with five states. The transition probabilities from state i to state j are denoted as a_{ij}. An observation o_t at time instant t, is created according to the output probability densities of state i, denoted as b_i.

5.4.6 Hidden markov model

Hidden Markov models (HMMs) are statistical models that have been utilized successfully, for example, in speech recognition [105]. A HMM consists of two stochastic processes. The first process is a Markov chain with N finite states. This process is not directly observable but is *hidden*, and it can be observed only through another stochastic process that produces the sequence of observations. A simple HMM structure with only left-to-right transitions between the five states is shown in figure 5.4. At each discrete time instant t, a decision for the next state of the Markov chain is made based on the transition probabilities from state i to state j, denoted as a_{ij}. After the next state is determined, an observation o_t is created according to the output probability densities of state i, denoted as b_i.

HMMs have been utilized in the envelope prediction of ABE methods [106, 78, 107]. In [78], each state of the HMM represents a typical highband spectral envelope. During the bandwidth extension, the narrowband signal frames are mapped onto the states of the HMM, and the parameters representing the highband envelope are determined using an estimation rule.

The narrowband feature vector in [78] consists of both parameters representing the envelope (LPC coefficients) and scalar features (such as the zero-crossing rate, normalized frame energy, gradient index, local kurtosis and spectral centroid) that contain voiced/unvoiced information. The highband envelope, on the other hand, is represented by cepstral coefficients.

The parameters of the Markov chain with states representing the spectral envelopes are obtained by vector quantizing the highband spectral

envelopes using a training data with true wideband speech. As a result, every state of the HMM corresponds to one entry of the VQ codebook. The resulting HMM structure is ergodic, indicating that a transition from a state to any other state is possible. Furthermore, for each state S_i, a statistical model is constructed based on the narrowband features $\mathbf{x}$ and the highband spectral envelopes $\mathbf{y}$. The statistical model consists of the state and transition probabilities $P(S_i)$ and $P(S_i(m+1)|S_j(m))$, respectively, the observation probability $p(\mathbf{x}|S_i)$, and the emission probability $p(\mathbf{y}|S_i)$. The observation probability, i.e. the PDF of the feature vectors $\mathbf{x}$, for each state $p(\mathbf{x}|S_i)$ is estimated by a GMM using the EM training procedure. Finally, the estimation rule defines how the coefficients representing the spectral envelope of the missing highband are formed. In [78], a minimum mean square error (MMSE) rule is applied, and the resulting highband envelope is the sum of individual codebook entries weighted by a posteriori probabilities of the corresponding states of the HMM.

An advantage of the algorithm presented in [78] is that, due to the HMM and the MMSE estimation rule, the algorithm is not memoryless, as most of the ABE methods, but information from the preceding frames is exploited. Also, the foundation of the algorithm is strongly based on a widely used and accepted mathematical model that has been successfully exploited in other fields of speech processing. On the other hand, the method is a "black box" implementation, and the performance is dependent on the feature selection and training.

5.4.7 Neural networks

Neural networks have been utilized in the estimation of the highband parameters, e.g. in [108, 109, 110, 69, 72]. A multilayer feedforward neural network (FFNN) is the most common neural network, and it was also used in [75, 109, 110, 69, 72]. It comprises an input layer of neurons, hidden layers, and an output layer of neurons. An example FFNN with one hidden layer is shown in figure 5.5. A hidden neuron v_j has m input signals x_k and an output v_j. Each input has a synaptic weight w_{jk}. An adder sums the weighted input signals and an optional bias, b_j, and finally the output of a neuron, v_j, is defined by an activation function $\varphi(\cdot)$ as follows:

$$v_j = \varphi(\sum_{j=1}^{m} w_{jk}x_k + b_j). \tag{5.18}$$

If every node in each layer is connected to every other node in the adja-

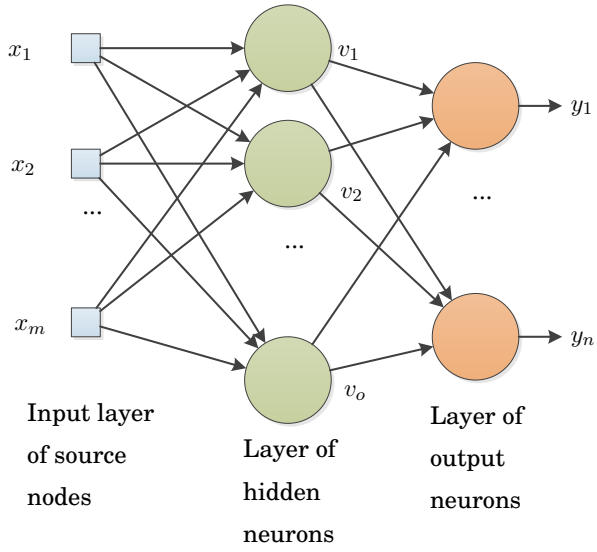

Figure 5.5. Feedforward neural network with one hidden layer between the input and output layers. The input layer comprises m input signals x_k, the hidden layer o hidden neurons v_j, and the output layer n output signals y_i.

cent forward layer, the neural network is said to be fully connected, otherwise it is partially connected. Furthermore, a basic FFNN can be modified by adding one or more feedback loops, i.e. by feeding output signals back to the inputs of the neurons. Such neural networks are called recurrent neural networks [111].

The parameters for a neural network are determined through a learning (or training) process, where the connection weights are adjusted iteratively. Usually the learning process is based on an example set of input data for which the network learns to produce the desired output and at the same time to generalize the results for other inputs. Learning methods are often divided into three groups: 1) supervised learning, 2) unsupervised learning, and 3) reinforcement learning. In supervised learning, the the desired behaviour of the neural network is described in terms of a set of input-output examples. An iterative learning process is implemented by minimizing an error function, i.e. a measure between the output of the network and the desired output, of the whole training data. In unsupervised learning, desired outputs are not available, and the learning is based on the correlations of the input data [112]. A typical use of unsupervised learning is clustering, where the system forms clusters of similarities from the input data. In reinforcement learning, the desired output of

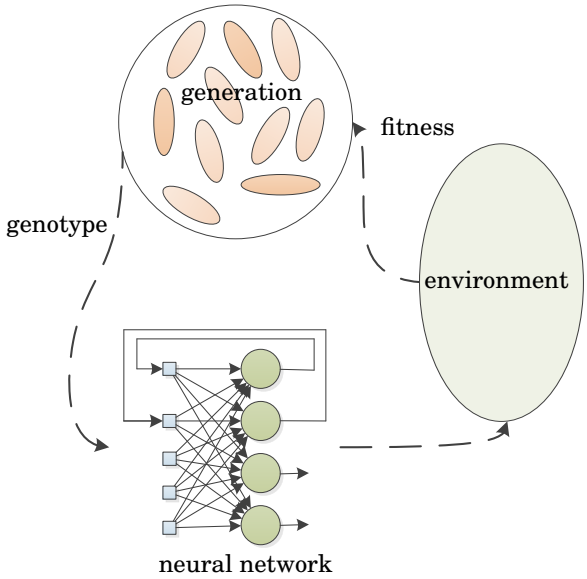

Figure 5.6. Schematic diagram of neuroevolution methods. Each genotype is tested by decoding it into a neural network and performing the task, resulting in a fitness value for the genotype. After evaluating all the genotypes in this manner, a new generation of genotypes is created by genetic operations among the best genotypes. This process is continued until the fitness is sufficiently high.

the network is also unknown, but the performance of the network can be measured by a so-called fitness function. The learning process is therefore reinforced by crediting a desired behaviour of the network and penalizing an undesired behaviour.

Neuroevolution methods are a special group of reinforcement learning methods that can be used not only for modifying neural network weights but also topologies [113]. They are also well suited for recurrent networks. The basic idea behind most of the neuroevolution methods is shown in figure 5.6. The first generation of the genotype population is first created. Each genotype is then tested by decoding it into a neural network and performing the task, resulting in a fitness value for the genotype. After all the genotypes are evaluated in this manner, a new generation of genotypes is created by genetic operations among the best genotypes. This process is continued until the fitness is sufficiently high.

A three-layer feed-forward neural network is used in [69] to determine weighting parameters of a folded narrowband spectrum in the highband (4-8 kHz), quite similarly to Publication II. Instead of utilizing neuroevolution learning, as in Publication II, a supervised learning method, called

the Levenberg-Marquardt algorithm, is used to train the network. The input to the neural network is a feature vector of nine narrowband features. The neural network is trained to output weights for critical bands in the highband. The shaping of the folded highband is implemented in the spectral domain by spline functions that are constructed around the critical band weights given by the neural network. The method was tested against both pure wideband and narrwoband references using a CCR test. The results showed clear preference towards ABE samples compared to narrowband, but wideband was also reported superior to ABE samples. Both naive and non-naive listeners were used in the test. In addition to the quality test, MRT was also utilized in evaluating intelligibility between narrowband and ABE samples. The reported results showed some improvement also in terms of intelligibility.

Another approach based on neural networks is introduced in [109]. The method closely follows the LPC-based algorithm model shown in figure 5.2. The mapping of the narrowband cepstral coefficients derived from the LPC coefficients onto wideband coefficients is implemented by a FFNN. A simple supervised learning rule called back-propagation is used. In addition, in order to use a neural network for envelope extension, an alternative implementation with linear mapping of the envelope parameters is presented, and the two methods are compared to each other. Both objective and subjective comparison tests reported in [109] indicate nearly the same quality improvement for both neural network and linear mapping-based ABE versions compared to a narrowband reference.

In [75], a neural network-based ABE algorithm was compared with a codebook-based method. The neural network topology was rather similar to the one presented in [109]. A multilayer perception (MLP) network with three layers and the back-propagation learning algorithm was used. Both input and output parameters for the neural network were LPC coefficients. The two methods were evaluated with both objective and subjective methods. Interestingly, the results from the objective and subjective tests were inconsistent; three of four objective measures indicated better performance for the neural network algorithm. However, the MOS test resulted in a clear preference for codebook mapping.

In [72], an ABE algorithm is studied that uses a neural network to estimate the mel spectrum for the highband. (This technique is essentially the same as the ABE2 algorithm in Publication VII.) In this method, the highband is divided into subbands that, in turn, are weighted to realize

the mel spectrum. The method uses a neuroevolution algorithm, called neuroevolution of augmenting topologies (NEAT), to train the neural network. The training starts from a minimum topology of the network that evolves during the training process. An advantage of the NEAT algorithm is that it can exploit recurrent connections and thus include memory within the network. According to the CCR listening test results reported in [72], the method improves narrowband-coded AMR speech significantly.

Despite their many advantages, neural networks also have disadvantages. In practice, there is an unlimited number of possible network topologies, training algorithms, and training data sets. Using a large set of training data is recommended, which easily leads to a slow development process of the algorithm. Both promising and unpromising results from evaluations of quality were reported in [108, 75, 109, 69, 72]. The comparisons between neural networks with linear mapping and codebooks [75, 109] indicate that with a simple neural network, the performance is perhaps not good enough, but with more complex structures and with perceptual fitness functions, it is possible to obtain good results [72].

6. Artificial bandwidth extension in mobile devices

Improved speech quality and intelligibility, achieved through a carefully designed and implemented ABE algorithm, can be exploited in products that receive or store narrowband speech. For example, cellular telephony, car telephony, cellular networks, VoIP telephony, and teleconferencing systems are potential products that benefit from ABE methods.

The quality experienced by the end user is dependent on the whole speech processing chain from the far-end user to the near-end user, including the acoustic properties of the user terminals. Therefore, when implementing an ABE method in a product, the properties of the present product should be taken into account. In addition, whether the product, and consequently the ABE method, is used in noisy places or not, influences the quality expectations of the ABE method in use. In a telephone conversation, intelligibility comes before audio quality, but after excellent intelligibility is achieved, the importance of quality increases. For example, in an extremely noisy environment, it is important to understand the message that the conversation partner is delivering and to be able to respond to the conversation. On the other hand, when speaking on the phone in a quiet place, the naturalness of the voice is valued by the end user, and being able to recognize the other person on the phone easily is appreciated.

In this section, the implementation of an ABE method in mobile devices and car telephony are discussed in more detail.

6.1 Artificial bandwidth extension in a mobile device

The fact that the most of the mobile phone users of the world's 5.6 billion mobile connections ([34]) are still offered only narrowband speech makes mobile devices an attractive target for an ABE method. During

the transition phase from narrowband to wideband speech communication, several scenarios for phone calls relative to the speech bandwidth exist, as illustrated in figure 6.1. If either of the devices in a telephone conversation support only narrowband coding, speech is transmitted in narrowband, as shown in figures 6.1 a–d. Only if both devices support wideband speech coding, true wideband speech is achieved, as illustrated in figure 6.1 e. ABE can be implemented in both narrowband and wideband terminals. In practice, in both cases wideband acoustical design is needed. However, whether the device is a narrowband or wideband terminal has an influence on type approval tests that are performed for each mobile device, following the 3GPP specifications [114]. For a narrowband mobile terminal, passing the narrowband performance requirements is sufficient, whereas a wideband terminal should meet wideband telephony performance requirements.

Several issues should be taken into account when implementing an ABE algorithm in a mobile device. The algorithm has to be suitable for real-time implementation with a reasonable computational load. The algorithmic delay should be small enough so that the overall downlink processing delay does not exceed 150 ms, because greater delays start to annoy the end users and make the conversation difficult.

There are thousands of languages in the world. The most spoken languages that are all spoken by over 100 million people include Chinese, Hindi, English, Spanish, Arabic, Portuguese, French, Bengali, Russian, Japanese, and German [115]. The languages differ from each other in temporal and spectral characteristics, and the ABE method should be as language independent as possible.

The algorithm should be speaker independent and result in good speech quality whether the speaker is male or female, an adult or a child. The far-end user might have a speaking disorder, a quiet or a loud voice. Speaker dependent methods have been reported to result in superior performance than speaker independent methods, at least in [78]. Perhaps in the future, the fact that a mobile device is often personal could provide an opportunity to develop an architecture where the algorithm adapts to special speech characteristics of the user.

Mobile devices are used everywhere in the world: in homes, in offices, in outdoor activities, in cars, in concerts, in discos, at construction sites, and so on. Therefore, the ABE method should be robust against speech signals with all kinds of background noise having both high and low SNR. The

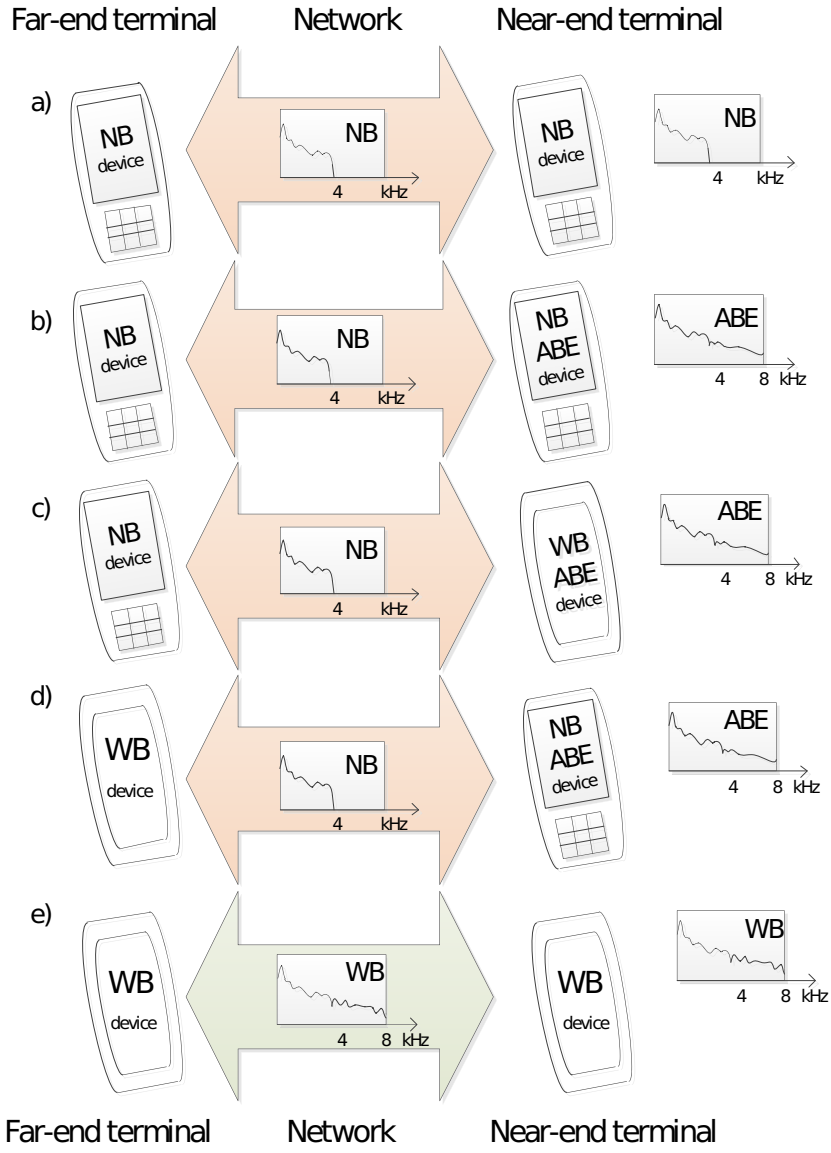

Figure 6.1. Different phone call scenarios between narrowband and wideband mobile devices. If both the far-end and the near-end terminals are narrowband devices, the speech signal is narrowband as well (a). True wideband is achieved only if both terminals and the network support wideband (e). Otherwise, the speech signal is narrowband, and ABE can be used to widen the spectrum (b, c, and d).

far-end user may be surrounded by any kind of background noise, which is included in the signal that serves as the input for the ABE method. The ABE method should be able to enhance the speech quality without increasing the noise level. Babble noise, cafeteria noise, and any other noise with high frequency content is extremely challenging for ABE methods because the methods easily start to enhance these high frequency components. Background noise around the end user masks possible artefacts introduced by an ABE method. Therefore, the ABE method is especially useful in mobile phones that are often used in noisy places.

Different speech codecs and bit rates with varying speech quality are used around the world. In addition to background noise, codec distortion deteriorates the quality of the narrowband input signal, and the performance of an ABE method should be assessed with distorted input signals of different kinds.

6.1.1 Signal path in mobile telephony

The voice quality in a mobile device is affected by the whole chain from the far-end user to the near-end user. The acoustic environment and the background noise conditions around the far-end user have an effect on the quality of the input signal that is captured by the far-end telephone device, which may be a handportable mobile device, a landline telephone, a car hands-free system, or a VoIP application. Taking a mobile device as an example, The microphone of the far-end user captures the speech signal. It is processed through a low-pass filter to remove unwanted signal components, and an analogue-to-digital conversion. The far-end signal path is then followed by speech enhancement algorithms such as noise suppression, echo cancellation, and level control that are implemented in the mobile device. The digital signal is then coded by a speech codec, like the AMR codec, and transmitted through a cellular network.

In the near-end device, the decoded signal is first encoded. The resulting digital signal is processed through speech enhancement algorithms such as noise suppression and dynamic level control. ABE processing is typically implemented after other speech enhancement, and the influence of the other algorithms on the ABE quality, especially noise suppression and dynamic range control (compression), should be assessed. After ABE processing, the frequency response of the speech signal is equalized for the transducer. Finally, there is a digital-to-analogue converter before the amplifier and the transducer.

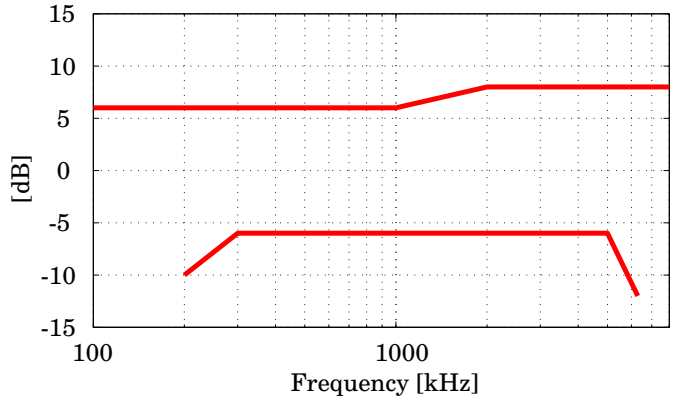

Figure 6.2. Receiving frequency mask as specified in [114].

6.1.2 Acoustic design of a mobile terminal

The acoustic design of a mobile terminal comprises the hardware components and their mounting, i.e. how they are placed in the device. The requirements for good ABE quality are basically the same as for good wideband speech, i.e. the earpiece should be able to reproduce a frequency band of 50–7000 Hz with reasonable earpiece distortion. Additionally, equalizers are required to compensate for the unideal frequency response of the hardware components, in particular the microphone and the loudspeaker. For mobile devices, ITU-T specifies frequency masks for both the receiving and sending quality that has to be met in the hand-portable mode of mobile devices [114]. There are specified requirements for both narrowband and wideband telephony. An ABE method can be implemented in a narrowband device or in a wideband device. In a narrowband device, only narrowband speech transmission is supported, and the ABE method is regarded as a narrowband speech enhancement feature. In a wideband device, wideband speech coding is supported and ABE is used if true wideband speech is not available.

The receiving wideband frequency mask, specified in the 3GPP specification 26.131 [114], is shown in figure 6.2. The frequency mask is relatively loose, which is why an ABE method should always be tested with the acoustics of the device. For example, if the ABE algorithm in use produces a spectral gap around 4 kHz, it might not be audible with a flat frequency response, but a non-flat response might boost the gap and thus increase audible distortion. The quality of the state-of-the-art ABE algorithms is not completely comparable with true wideband speech. There are some

artefacts and distortion in the extended signals, but an adequate ABE quality in a product can usually be achieved by finding a good balance between the original narrowband signal and the regenerated highband one.

6.2 Artificial bandwidth extension in car telephony

In many countries, the use of the hands-free cellular phone is mandatory while driving a car for road safety reasons. Hands-free cellular phone usage improves road safety since the driver can better concentrate on driving without having to hold the mobile device to the ear with his or her hand, or even with a shoulder, if both hands are needed for some driving manoeuvre. Hands-free usage of a mobile phone in a car can be achieved by, for example, utilizing the speaker mode of a mobile device, using a wired headphone, through a bluetooth accessory, or by utilizing an integrated in-car hands-free system.

Car telephony is an attractive target for ABE methods, because background noise degrades speech intelligibility and increases listening fatigue. In addition, the car interior, which typically has noise with low-pass characteristics, masks possible artefacts generated by ABE processing. In [116], the usage of a bandwidth extension algorithm improved intelligibility of narrowband speech of meaningless vowel-consonant-vowel syllables in a car environment. Some quality improvement was also reported in low (0 dB) SNR conditions. Furthermore, when utilizing ABE in car telephony through integrated hands-free systems (for example, [117]), the acoustic environment around the near-end speaker is always known, and the whole signal processing path can be designed for car usage, such as in [118].

7. Summary of the publications

This thesis consists of seven articles four of which were published in international reviewed journals and three in full-paper reviewed conferences. All the articles focus on ABE of speech signals.

Publication I: "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech"

The first publication is a conference article that presents an ABE algorithm. The method is based on spectral folding, i.e., the narrowband spectrum from the frequency range 0–4 kHz is folded onto the highband (4–8 kHz). The folded spectrum is shaped in the FFT domain by spline functions that are optimized using a genetic algorithm. The performance of the algorithm is evaluated by an intelligibility test called SRT and by an ACR test that is often used to assess speech codecs. The results from the SRT test indicate that the algorithm improves speech intelligibility in noise. The ACR test results show no statistical improvement in speech quality compared to narrowband AMR speech. The performance of the algorithm is further evaluated in Publication III.

Publication II: "Neural network-based artificial bandwidth expansion of speech"

An ABE algorithm utilizing a neural network is presented in the article. The algorithm, called neuroevolution ABE (NEABE), is based on spectral folding. A feature vector is computed from the narrowband signal and serves as an input to a neural network. As an output, the neural network provides parameters required to formulate a spline function that, in turn, is used to shape the highband spectrum. A neuroevolution algorithm is

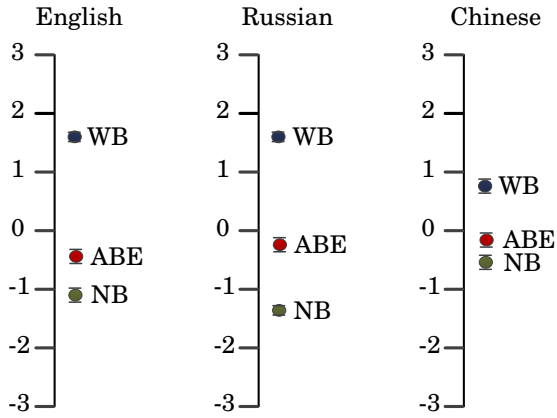

Figure 7.1. Main results of the CCR language test presented in the order of preference — wideband (WB), ABE, and narrowband (NB) processings — in three languages. Mean scores are indicated by dots and 95 % confidence intervals are shown with error bars.

used to train the neural network.

The algorithm is evaluated against a clean narrowband reference by a paired comparison test and with a CCR test. The tests indicate a clear preference towards NEABE-processed speech compared to the clean narrowband reference. In addition, an objective measure, the SD, between NEABE outputs and true wideband signals are computed for phonetically labelled speech data.

Publication III: "Evaluation of an artificial speech bandwidth extension method in three languages"

In this journal article, a potential language dependency of an ABE algorithm is studied. The algorithm is an enhanced version of the method presented in Publication I, and it is evaluated against an AMR-coded narrowband reference and an AMR-WB-coded wideband reference. Three CCR tests are arranged in English, Russian and Mandarin Chinese. All the languages selected for the evaluation are spoken by millions of people around the world. Mandarin Chinese is chosen as one of the test languages because it is a tonal language. Furthermore, Russian is chosen because of its varied set of fricative sounds that are known to be challenging for ABE methods. The test results are consistent in all three languages, indicating that the algorithm is not language dependent. The main results of three CCR tests are shown in figure 7.1.

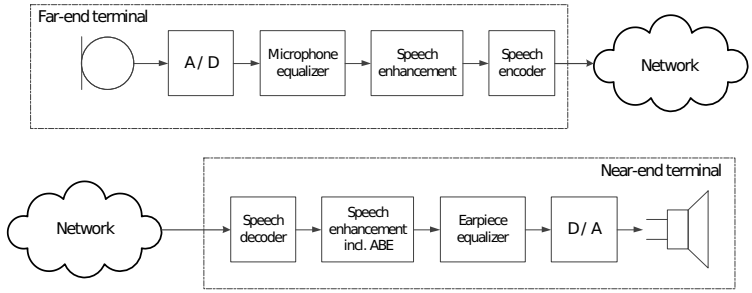

Figure 7.2. Signal path from the far-end user to the near-end user in a telephone conversation between two mobile phone users.

Publication IV: "Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal"

This article discusses the development cycle of an ABE method from an initial idea to the implementation in a mobile terminal. In addition to the algorithm development, the process includes several subjective tests and simulations that verify the proper performance of the algorithm in different use scenarios.

The article also discusses the utilization of the ABE technology in a mobile device. The signal path from the far-end user to the near-end is illustrated in figure 7.2. The sending and receiving mobile terminals have certain characteristics including the influence of speech coding, other speech enhancement algorithms, and acoustical design of the devices. In addition, background noise at both ends of the communication affects the speech quality experienced by the end user.

Publication V: "Binaural artificial bandwidth extension (B-ABE) for speech"

In this conference article, the requirements for the ABE algorithm to extend binaural signals are discussed. An ABE algorithm originally designed for monaural signals is modified for binaural speech signals so that the algorithm preserves the localization information of the speakers. A subjective listening test is also organized to analyse the performance of the method. The test results indicate that the algorithm preserves the localization information well.

Binaural speech technology can benefit from spatial audio, and if wide-

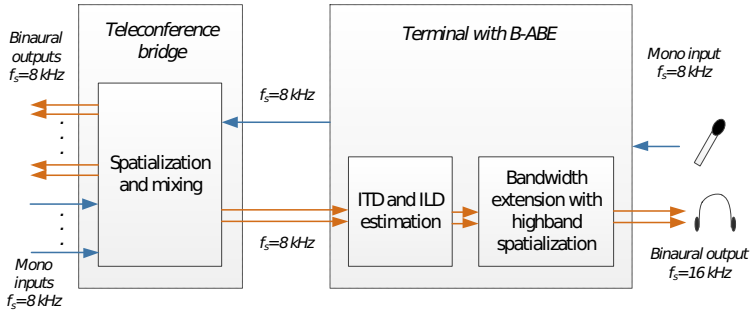

Figure 7.3. Teleconferencing system including a conference bridge and a terminal with the B-ABE function. 3D processing is performed in the conference bridge and the binaural signal is sent to the terminal where B-ABE processing takes place.

band speech is not available, ABE can be exploited. An example teleconferencing system with 3D processing implemented in a conference bridge and B-ABE in a terminal is shown in figure 7.3.

Publication VI: "Evaluating artificial bandwidth extension by conversational tests in car using mobile devices with integrated hands-free functionality"

This article reports conversational test results for ABE processing. In the test, phone calls between two persons are made using mobile devices over a cellular network. Three connection types are involved in the test: a narrowband connection with AMR coding, a wideband connection with AMR-WB coding, and an ABE connection with AMR coding and ABE processing implemented in the terminal. The test subjects held short conversations in a car with and without simulated car interior noise. The subject in the car was able to switch between two different connections during each phone call, and after the conversation, the subject was asked which connection was better. The placement of the mobile devices in the car is shown in figure 7.4.

The results show a clear preference for the ABE connection compared to the narrowband one, whereas wideband was preferred over both narrowband and ABE. To the best of the author's knowledge, this article is the first to report conversational testing of an ABE method. In addition, for the first time, an ABE algorithm is evaluated with an ABE method running on a true end product and using real cellular networks.

Figure 7.4. Mobile devices in the test car.

Publication VII: "Conversational quality evaluation of artificial bandwidth extension of telephone speech"

The last article discusses conversational quality testing of ABE more thoroughly. An extensive conversational evaluation consisting of two separate tests was organized. Two different ABE algorithms, a narrowband reference (AMR), and a wideband reference (AMR-WB) were included in the tests. The first test closely followed the ITU-T P.805 recommendation, whereas the second test was a modification of the first and enabled a paired comparison and asymmetric conversation tasks. A schematic illustration of the test setup is shown in figure 7.5.

The results indicate that speech processed with one of the ABE methods is preferred to narrowband speech in noise. However, wideband speech is superior to both narrowband and ABE-processed speech.

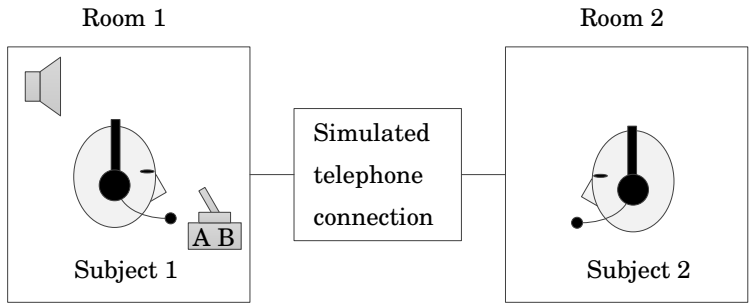

Figure 7.5. Schematic illustration of the conversation test setup.

8. Conclusions

Narrowband speech coding that is deployed in many telephone applications limits the speech bandwidth to the frequency range 300–3400 Hz, resulting in degraded speech quality and intelligibility. When only narrowband speech transmission is available, ABE can be applied to the signal at the receiving end of the communication link. An ABE method aims to improve speech quality and intelligibility by regenerating new frequency components in the signal in frequencies above 3400 Hz. Since the extension is made without any transmitted information, the algorithm is suitable for any telephone application that is able to reproduce wideband audio signals.

In this thesis, ABE towards high frequencies is addressed from the mobile communication perspective. The following issues are addressed:

1. ABE algorithm development for narrowband speech signals.

2. ABE for binaural signals.

3. Evaluation of the ABE methods through subjective listening tests and conversational tests.

4. Implementation of an ABE method in a mobile device.

The developed ABE algorithms are presented in Publication I (enhanced algorithms in Publication III and Publication VII), Publication II and Publication VII. All the three methods are suitable for real-time implementation with a reasonable computational load and delay due to the algorithm. The methods are primarily designed for monaural telephone speech signals, whereas the extension of binaural signals is addressed in Publica-

tion V.

For evaluating the developed ABE methods, several listening opinion tests have been conducted. The quality of ABE-processed signals are compared with both narrowband and wideband references through subjective listening tests in laboratory conditions. The selected test methods include ACR (in Publication I and Publication IV), CCR (in Publication II and Publication III) and paired comparison (Publication II). In addition, the CCR test reported in Publication III addresses language dependency issues of an ABE method. Furthermore, SRT in noise test was utilized to assess how ABE affects speech intelligibility in noise (in Publication I and Publication IV). In Publication VI and Publication VII, the ABE quality is assessed by conversational tests, where the algorithms are evaluated in a more natural speech communication situation between mobile phone users. The implementation of the ABE algorithm presented in Publication I, Publication III, and Publication IV in a mobile phone is discussed in Publication IV.

A thorough subjective evaluation verifies that narrowband speech quality and intelligibility can be improved by the novel ABE algorithms developed in this thesis. The results from the tests are consistent and indicate no language dependency. Furthermore, the results have been obtained with realistic speech data, i.e. coded speech from several speakers and languages have been involved in the tests.

The results from the SRT test and the conversational test reported in Publication VI indicate that ABE improves quality and intelligibility especially in a noisy environment where the artefacts are not heard due to the masking effect. However, the larger scale conversational test in Publication VII does not completely verify this result, even though the results indicate that ambient noise increased the effort needed to understand the conversation partner, and one of the ABE methods reduced the effort to understand female voices compared with narrowband AMR.

ABE for binaural signals has to be implemented such that the binaural cues are preserved. Especially, the ILD is an important cue at high frequencies. In Publication V, the ABE processing improved the 3D localization compared to the narrowband signal.

In a mobile phone, the performance of an ABE method depends on the entire processing chain of the downlink speech signal. If the equalized frequency response of the earpiece is not sufficiently flat, some artefacts of the ABE processing may be emphasized. A possibility to tune the algo-

rithm for each acoustical design separately is beneficial. If the acoustical properties of a mobile device are not optimal for an ABE method, uplink noise dependent tuning can be utilized.

Three different ABE algorithms have been developed and evaluated in this thesis. The results from the tests are consistent, indicating that even though the extension is completely artificial, speech quality and intelligibility is primarily improved. However, there is still a gap between the quality of ABE processed speech and true wideband speech. In the future, it would be interesting to study if the quality gap could be further decreased by bandwidth extension for low frequencies. Moreover, the small-scaled conversational test in Publication VI is the only test to date where ABE was evaluated using true mobile devices and cellular connections. The performance of the algorithm was carefully tuned for the particular acoustical design. Despite the fact that the results are promising, even more in-depth field studies would be beneficial in finding out whether the end users adapt to the artefacts introduced by the ABE method or whether they start to annoy the users. In addition, based on this study, an open research question is how telephone users experience the hand-overs for narrowband to wideband speech and vice versa, compared to hand-overs from wideband to ABE and vice versa.

Bibliography

[1] International Telecommunication Union. ITU-T Recommendation G.712, Transmission performance characteristics of pulse code modulation channels, 2001.

[2] B. C. J. Moore and C.-T. Tan. Perceived naturalness of spectrally distorted speech and music. *The Journal of Acoustical Society of America*, 114(1):408–419, 2003.

[3] H. Fletcher and R. H. Galt. The perception of speech and its relation to telephony. *The Journal of Acoustical Society of America*, 22(2):89–151, 1950.

[4] The European Telecommunications Standards Institute (ETSI). Digital cellular telecommunications system (Phase 2); Enhanced Full Rate (EFR) speech processing functions; General description (GSM 06.51 version 4.1.1), 2000.

[5] 3rd Generation Partnership Project (3GPP). Adaptive multi-rate (AMR) speech codec, transcoding functions, 3GPP TS 26.090, version 10.0.0., 2011.

[6] 3rd Generation Partnership Project (3GPP). Adaptive multi-rate - wideband (AMR-WB) speech codec, transcoding functions, 3GPP TS 26.190, version 6.1.1., 2005.

[7] Orange launches world's first high-definition voice service for mobile phones in Moldova. `http://event.orange.com/media/hd_voice_en/UPL8498801975512936614_CP_Orange_Moldova_HDVoice_en.pdf`, 2009. [accessed 26-July-2012].

[8] SILK: Super wideband audio codec. `https://developer.skype.com/silk`. [accessed 17-June-2012].

[9] B. Geiser, P. Jax, and P. Vary. Artificial bandwidth extension of speech supported by watermark-transmitted side information. In *Proceedings of Interspeech*, page 1497–1500, Lisboa, Portugal, 2005.

[10] A. Sagi and D. Malah. Bandwidth extension of telephone speech aided by data embedding. *EURASIP Journal on Advances in Signal Processing*, 2007(1), 2007.

[11] S. Chen and H. Leung. Speech bandwidth extension by data hiding and phonetic classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 593–596, Honolulu, Hawaii, USA, 2007.

[12] D. L. Richards. *Telecommunication by Speech, The Transmission Performance of Telephone Networks*. London, Butterworths, 1973.

[13] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, New Jersey, 1978.

[14] J. L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, 1965.

[15] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve. The effect of high-band harmonic structure in the artificial bandwidth expansion of telephone speech. In *Proceedings of Interspeech*, pages 2497–2500, Antwerp, Belgium, 2007.

[16] G. Fant. *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[17] J. Makhoul. Linear prediction: a tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[18] T. D. Rossing. *The Science of Sound*. Addison Wesley, 1990.

[19] J. Blauert, editor. *Communication Acoustics*. Springer-Verlag, 2005.

[20] M. Karjalainen. *Kommunikaatioakustiikka*. Teknillinen korkeakoulu, 1999.

[21] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 2001.

[22] O. Lodge. The history and development of the telephone. *Journal of the Institution of Electrical Engineers*, 64(359):1098–1114, 1926.

[23] S. Voran. Listener ratings of speech passbands. In *Proceedings of the IEEE Workshop on Speech Coding For Telecommunications*, pages 81–82, 1997.

[24] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19(1):90–119, 1947.

[25] E. Larsen and R. M. Aarts. *Audio Bandwidth Extension - Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. Wiley, 2004.

[26] Global mobile suppliers association (GSA). Mobile HD voice: Global update report, GSM/3G market/technology update, May 21, 2012. `http://www.gsacom.com/cgi/redir.pl5?url=http://www.gsacom.com/downloads/pdf/GSA_Mobile_HD_Voice_report_210512_hdvz.php4`, 2012. [accessed 19-August-2012].

[27] R. V. Cox, S. F. de Campos Neto, C. Lamblin, and M. H. Sherif. ITU-T coders for wideband, superwideband, and fullband speech communication. *IEEE Communications Magazine*, 47(10):106–109, 2009.

[28] J. D. Gibson. Speech coding methods, standards, and applications. *IEEE Circuits and Systems Magazine*, 5(4):30–49, 2005.

[29] International Telecommunication Union. ITU-T Recommendation G.114, One-way transmission time, 2007.

[30] F. Hammer, P. Reichl, and A. Raake. The well-tempered conversation: interactivity, delay and perceptual voip quality. In *Proceedings of the IEEE International Conference on Communications (ICC)*, pages 244–249, Seoul, Korea, 2005.

[31] International Telecommunication Union. ITU-T Recommendation G.711, Pulse code modulation (PCM) of voice frequencies, 1972.

[32] International Telecommunication Union. ITU-T Recommendation G.726, 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM), 1990.

[33] ETSI. Cellular history. `http://etsi.org/WebSite/Technologies/Cellularhistory.aspx`. [accessed 10-June-2012].

[34] Gartner. Gartner says worldwide mobile connections will reach 5.6 billion in 2011 as mobile data services revenue totals $ 314.7 billion. `http://www.gartner.com/it/page.jsp?id=1759714`, 2011. [accessed 10-June-2012].

[35] A. S. Spanias. Speech coding: A tutorial review. *Proceedings of the IEEE*, 82(10):1541–1582, 1994.

[36] A. M. Kondoz. *Digital Speech: Coding for Low Bit Rate Communication Systems*. Wiley, 2004.

[37] P. Ojala, A. Lakaniemi, H. Lepänaho, and M. Jokimies. The adaptive multirate wideband speech codec: system characteristics, quality advances, and deployment strategies. *IEEE Communications Magazine*, 44(5):59–65, 2006.

[38] M. Poikselkä, H. Holma, J. Hongisto, J. Kallio, and A. Toskala. *Voice over LTE (VoLTE)*. Wiley, 2012.

[39] K. Järvinen, I. Bouazizi, L. Laaksonen, P. Ojala, and A. Rämö. Media coding for the next generation mobile system LTE. *Computer Communications*, 33(16):1916–1927, 2010.

[40] 3GPP TR 26.935. Packet switched (PS) conversational multimedia applications; performance characterisation of default codecs, 2011.

[41] A. Raake. *Speech Quality of VoIP: Assessment and Prediction*. Wiley, 2006.

[42] J. Blauert and U. Jekosch. Sound-quality evaluation - a multi-layered problem. *Acta Acustica*, 83(5):747–753, 1997.

[43] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac. On the evaluation of the conversational speech quality in telecommunications. *EURASIP Journal on Advances in Signal Processing*, 2008, 2008.

[44] P. Loizou. *Multimedia Analysis, Processing and Communications*, chapter Speech quality assessment. Springer Verlag, 2011.

[45] S. Möller and A. Raake. Telephone speech quality prediction: Towards network planning and monitoring models for modern network scenarios. *Speech Communication*, 38(1-2):47–75, 2002.

[46] T. Letowski. Sound quality assessment: Concepts and criteria. In *Proceedings of the 87th Audio Engineering Society (AES) Convention (Paper D-8, Preprint 2825)*.

[47] N. Côté. *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Springer, 2011.

[48] International Telecommunication Union. ITU-T Recommendation P.835, Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, 2003.

[49] International Telecommunication Union. ITU-T Recommendation P.800, Methods for subjective determination of transmission quality, 1996.

[50] International Telecommunication Union. ITU-T recommendation P.805, Subjective evaluation of conversational quality, 2007.

[51] International Telecommunication Union. ITU–T Recommendation G.107, The E-model, a computational model for use in transmission planning, 2008.

[52] International Telecommunication Union. ITU-T Recommendation 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, 2001.

[53] International Telecommunication Union. ITU-T Recommendation P.862.2, Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs, 2005.

[54] International Telecommunication Union. ITU-T Recommendation P.863, Perceptual objective listening quality assessment, 2011.

[55] G. Fairbanks. Test of phonemic differentiation: The rhyme test. *The Journal of Acoustical Society of America*, 30(7):596–600, 1958.

[56] W. D. Voiers. Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1:30–39, 1983.

[57] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter. Articulation-testing methods: Consonantal differentiation with a closed response set. *The Journal of Acoustical Society of America*, 37(1):158–166, 1965.

[58] International Telecommunication Union. ITU-T Recommendation G.729, Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP), 2007.

[59] U. Kornagel. Techniques for artificial bandwidth extension of telephone speech. *Signal Processing*, 86(6):1296–1306, 2006.

[60] H. Gustafsson and U. A. Lindgren amd I. Claesson. Low-complexity feature-mapped speech bandwidth extension. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):577–588, 2006.

[61] H. Pulakka, U. Remes, S. Yrttiaho, K. J. Palomäki, M. Kurimo, and P. Alku. Low-frequency bandwidth extension of telephone speech using sinusoidal synthesis and gaussian mixture model. In *Proceedings of Interspeech*, volume 1, pages 1181–1184, Florence, Italy, 2011.

[62] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaumé, 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*, 15(8):2496–2509, 2007.

[63] M. Nilsson, H. Gustafsson, S. V. Andersen, and B. Kleijn. Gaussian mixture model based mutual information estimation between frequency bands in speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 525–528, Orlando, FL, USA, 2002.

[64] P. Jax and P. Vary. An upper bound on the quality of artificial bandwidth extension of narrowband speech signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 237–240, Orlando, FL, USA, 2002.

[65] Y. Agiomyrgiannakis and Y. Stylianou. Combined estimation/coding of highband spectral envelopes for speech spectrum expansion. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 469–472, Montreal, Canada, 2004.

[66] A. H. Nour-Eldin, T. Z. Shabestary, and P. Kabal. The effect of memory inclusion on mutual information between speech frequency bands. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 53–56, Toulouse, France, 2006.

[67] A. H. Nour-Eldin and P. Kabal. Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech. In *Proceedings of Interspeech*, pages 2489–2492, Antwerp, Belgium, 2007.

[68] H. Yasukawa. Signal restoration of broad band speech using nonlinear processing. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 987–990, Trieste, Italy, 1996.

[69] T. V. Pham, F. Schaefer, and G. Kubin. A novel implementation of the spectral shaping approach for artificial bandwidth extension. In *Proceedings of IEEE International Conference on Communications and Electronics (ICCE)*, pages 262–267, Nha Trang, Vietnam, 2010.

[70] H. Carl and U. Heute. Bandwidth enhancement of narrow-band speech signals. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, volume 2, pages 1178–1181, Edinburgh, Scotland, 1994.

[71] P. Jax and P. Vary. Wideband extension of telephone speech using a hidden markov model. In *Proceedings of the IEEE Workshop on Speech Coding*, pages 133–135, Delavan, WI, USA, 2000.

[72] 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*, 19(7):2170–2183, 2011.

[73] Y. Nakatoh, M. Tsushima, and T. Norimatsu. Generation of broadband speech from narrowband speech using piecewise linear mapping. In *Proceedings of the European Conference on Speech Communication and Tech-*

nology (EUROSPEECH), volume 3, pages 1643–1646, Rhodes, Greece, 1997.

[74] I. Y. Soon, S. N. Koh, C.K. Yeo, and W. H. Ngo. Transformation of narrowband speech into wideband speech with aid of zero crossings rate. *Electronics Letters*, 38(24):1607–1608, 2002.

[75] B. Iser and G. Schmidt. Neural networks versus codebooks in an application for bandwidth extension of speech signals. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 565–568, Geneva, Switzerland, 2003.

[76] T. Ramabadran and M. Jasiuk. Artificial bandwidth extension of narrowband speech signals via high-band energy estimation. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, 2008.

[77] K.-T. Kim, M.-K. Lee, and H.-G. Kang. Speech bandwidth extension using temporal envelope modeling. *IEEE Signal Processing Letters*, 15:429–432, 2008.

[78] P. Jax and P. Vary. On artificial bandwidth extension of telephone speech. *Signal Processing*, 83(8):1707–1719, 2003.

[79] J. Makhoul and M. Berouti. High-frequency regeneration in speech coding systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 428–431, 1979.

[80] H. Yasukawa. Wideband speech recovery from bandlimited speech in telephone communications. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 4, pages 202–204, 1998.

[81] N. Enbom and W. B. Kleijn. Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients. In *Proceedings of the IEEE Workshop on Speech Coding*, pages 171–173, Porvoo, Finland, 1999.

[82] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter. Speech enhancement via frequency bandwidth extension using line spectral frequencies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 665–668, Salt Lake City, USA, 2001.

[83] C. Yagh and E. Erzin. Artificial bandwidth extension of spectral envelope with temporal clustering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5096–5099, Prague, Czech Republic, 2011.

[84] P. Jax and P. Vary. Enhancement of band-limited speech signals. In *Proceedings of Aachen Symposium on Signal Theory (ASST)*, pages 331–336, Aachen, Germany, 2001.

[85] U. Kornagel. Spectral widening of the excitation signal for telephone-band speech enhancement. In *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 215–218, Darmstadt, Germany, 2001.

[86] C.-F. Chan and W.-K. Hui. Wideband re-synthesis of narrowband CELP-coded speech using multiband excitation model. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP)*, volume 1, pages 322–325, Philadelphia, PA , USA, 1996.

[87] D. G. Raza and C. F. Chan. Quality enhancement of CELP coded speech by using an MFCC based gaussian mixture model. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 541–544, Geneva, Switzerland, 2003.

[88] B. Iser and G. Schmidt. Bandwidth extension of telephony speech. *EURASIP Newsletters*, 16(2):2–24, 2005.

[89] J. Epps. *Wideband Extension of Narrowband Speech for Enhancement and Coding*. PhD thesis, School of Electrical Engineering and Telecommunications, The University of New South Wales, 2000.

[90] Y. Qian and P. Kabal. Dual-mode wideband speech recovery from narrowband speech. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1433–1436, Geneva, Switzerland, 2003.

[91] R. Hu, V. Krishnan, and D. V. Anderson. Speech bandwidth extension by improved codebook mapping towards increased phonetic classification. In *Proceedings of Interspeech*, pages 1501–1504, Lisbon, Portugal, 2005.

[92] T. Unno and A. McCree. A robust narrowband to wideband extension system featuring enhanced codebook mapping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 805–808, Philadelphia, PA, USA, 2005.

[93] F. Itakura. Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57(S1):35, 1975.

[94] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[95] J. W. Paulus. Variable bitrate wideband speech coding using perceptually motivated thresholds. In *Proceedings of the IEEE Workshop on Speech Coding for Telecommunications*, pages 35–36, 1995.

[96] P. Jax. *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*. PhD thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 2002.

[97] Y. Yoshida and M. Abe. An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping. In *Proceedings of the Third International Conference on Spoken Language Processing (ICSLP)*, pages 1591–1594, Yokohama, Japan, 1994.

[98] C.-F. Chan and W.-K. Hui. Quality enhancement of narrowband CELP-coded speech via wideband harmonic re-synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1187–1190, Munich, Germany, 1997.

[99] J. Epps and W. H. Holmes. A new technique for wideband enhancement of coded narrowband speech. In *Proceedings of the IEEE Workshop on Speech Coding*, pages 174–176, Porvoo, Finland, 1999.

[100] Y. Qian and P. Kabal. Wideband speech recovery from narrowband speech using classified codebook mapping. In *Proceedings of Australian International Conference on Speech Science and Technology*, pages 106–111, Melbourne, Australia, 2002.

[101] K.-Y. Park and H. S. Kim. Narrowband to wideband conversion of speech using GMM based transformation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1843–1846, Istanbul, Turkey, 2000.

[102] A. H. Nour-Eldin and P. Kabal. Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech. In *Proceedings of Interspeech*, pages 53–56, Brisbane, Australia, 2008.

[103] A. H. Nour-Eldin and P. Kabal. Combining frontend-based memory with MFCC features for bandwidth extension of narrowband speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4001–4004, Taipei, Taiwan, 2009.

[104] A. H. Nour-Eldin and P. Kabal. Memory-based approximation of the gaussian model framework for bandwidth extension of narrowband speech. In *Proceedings of Interspeech*, volume 1, pages 1185–1188, Florence, Italy, 2011.

[105] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[106] M. Hosoki, T. Nagai, and A. Kurematsu. Speech signal band width extension and noise removal using subband HMM. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 245–248, Orlando, FL, USA, 2002.

[107] G. Chen and V. Parsa. HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 709–712, Montreal, Canada, 2004.

[108] A. Uncini, F. Gobbi, and F. Piazza. Frequency recovery of narrow-band speech using adaptive spline neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 997–1000, Phoenix, AZ, USA, 1999.

[109] D. Zaykovskiy and B. Iser. Comparison of neural networks and linear mapping in an application for bandwidth extension. In *10th International Conference on Speech and Computer (SPECOM)*, Patras, Greece, 2005.

[110] A. Shahina and B. Yegnanarayana. Mapping neural networks for bandwidth extension of narrowband speech. In *Proceedings of Interspeech*, pages 1435–1438, Pittsburgh, USA, 2006.

[111] S. Haykin. *Neural Networks - A Comprehensive Foundation*. Prentice Hall, 1999.

[112] X. Yao. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447, 1999.

[113] R. Miikkulainen. Encyclopedia of machine learning, Neuroevolution, New York, Springer. http://nn.cs.utexas.edu/?miikkulainen:encyclopedia10-ne, 2010. [accessed 10-June-2012].

[114] 3GPP TS 26.131. Terminal acoustic characteristics for telephony; requirements, 2011.

[115] One World Nations Online. Most widely spoken languages in the world. http://www.nationsonline.org/oneworld/most_spoken_languages.htm, 2012. [accessed 10-June-2012].

[116] P. Bauer, M.-A. Jung, Q. Junge, and T. Fingscheidt. On improving speech intelligibility in automotive hands-free systems. In *IEEE International Symposium on Consumer Electronics (ISCE)*, Braunschweig, Germany, 2010.

[117] QNX aviage acoustic processing suite. http://www.qnx.com/products/acoustic/index.html, 2012. [accessed 10-June-2012].

[118] B. Iser and G. Schmidt. Receive side processing for automotive hands-free systems. In *Proceedings of the Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pages 236–239, Trento, Italy, 2008.

Errata

Publication VII

In section V, "The frequency response of a speaker phone microphone in a mobile device differs from the frequency response of high quality headphones", should be: "The frequency response of a speaker phone loudspeaker in a mobile device differs from the frequency response of high quality headphones".

Telephone conversation in noisy environments is often difficult. Both speech intelligibility and quality are degraded by ambient noise. Furthermore, most of the mobile phone users are provided even today only with a limited range (300-3400 Hz) of voice frequencies due to the narrowband speech coding in cellular networks. This bandwidth is much narrower than what is necessary for high speech quality. Therefore, cellular operators are gradually starting to support wideband (50-7000 Hz) speech. However, during the transition from the existing narrowband systems to true wideband transmission narrowband speech can be enhanced by *artificial bandwidth extension* (ABE). In this thesis, ABE methods that enhance speech quality by adding high frequency components to the narrowband speech signal in the receiving mobile device are studied. The results indicate that ABE methods ease speech communication, especially in ambient noise, by providing the user of the mobile device speech of wider bandwidth than what was transmitted through the cellular network.

Aalto-DD 64/2013

ISBN 978-952-60-5124-6
ISBN 978-952-60-5125-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS