Techniques for speech intelligibility enhancement in mobile telephony

Emma Jokinen
Techniques for speech intelligibility enhancement in mobile telephony

Emma Jokinen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall AS1 of the school on 8 December 2017 at 12.00.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
Supervising professor
Academy Professor Paavo Alku, Aalto University, Finland

Preliminary examiners
Professor Martin Cooke, Universidad del País Vasco, Spain
Professor Tim Fingscheidt, Technische Universität Braunschweig, Germany

Opponent
Professor Gernot Kubin, Graz University of Technology, Austria
Abstract

Today’s consumers can use their mobile telephony devices almost anywhere and at any time. This means that speech communication is often disturbed by environmental background noise, making it hard for the listener to understand what the speaker is saying. To further aggravate the situation, the listener and the speaker are typically in different locations when the communication is taking place. This means that without listener feedback the speaker is unable to adjust his or her speaking style to fit the listening environment, as is normally done in face-to-face communication situations. However, speech communication by mobile telephony in noisy conditions can be improved using intelligibility enhancement technology.

This thesis contributes to the development of intelligibility enhancement techniques that can in principle be applied in real-time speech communication in a mobile device. The algorithms are intended to be used in a post-processing block in the receiving device to combat near-end noise in the listener’s environment. The target application places tight restrictions on the algorithmic delay, which means that frame-based processing in short time frames (for instance, 10 to 20 ms in length) must be employed.

Several algorithms for intelligibility improvement are proposed and their performance is demonstrated with subjective tests using simulated telephone speech. The majority of the introduced algorithms aim to mimic modifications that human speakers naturally employ when talking in noisy situations. In addition, a feature extraction technique that can be used to estimate the spectral tilt caused by the glottal excitation from telephone speech is proposed. Finally, the impact of noisy far-end conditions on post-processing in the receiving device is investigated. In general, the proposed post-processing techniques show clear intelligibility improvement over unprocessed telephone speech, ranging up to a 40 percentage point reduction in word-error rates.

Keywords  speech intelligibility enhancement, telephone speech, near-end noise, human speech production

ISSN-L 1799-4934
ISSN (printed) 1799-4934  ISSN (pdf) 1799-4942
Location of publisher Helsinki  Location of printing Helsinki  Year 2017
Tekijä
Emma Jokinen

Väitöskirjan nimi
Tekniikko puheen ymmärrättävyyden parantamiseen mobiilikaitteissa

Julkaisija
Sähköteknian korkeakoulu

Yksikkö
Signaalinkäsittelyn ja akustiikan laitos

Sarja
Aalto University publication series DOCTORAL DISSERTATIONS 208/2017

Tutkimusala
Puhe- ja kieliteknologia

Käsikirjoituksen pvm
28.07.2017

Julkaisuluvan myöntämispäivä
03.10.2017

Kieli
Englanti

Monografia
Artikkeliväitöskirja

Avainsanat
puheen ymmärrättävyyden parannus, puhelinpuhe, taustamelu, ihmisen puheentuotto

ISBN (painettu) 978-952-60-7679-9

ISSN-L 1799-4934

Julkaisupaikka Helsinki

Sivumäärä 159


Avainsanat
puheen ymmärrättävyyden parannus, puhelinpuhe, taustamelu, ihmisen puheentuotto

ISBN (pdf) 978-952-60-7684-3

ISSN (painettu) 1799-4934

ISSN (pdf) 1799-4942

Julkaisupaikka Helsinki

Painopaikka Helsinki

Vuosi 2017

Sivumäärä 159


This work was carried out in the Department of Signal Processing and Acoustics at Aalto University in Espoo, Finland. The work was funded by Nokia, the Academy of Finland, and the Finnish Graduate School of Electronics, Telecommunications, and Automation (GETA). Financial support in the form of personal grants was also received from the Nokia foundation, the HPY research foundation, Tekniikan Edistämissäätiö (TES), and the Walter Ahlström foundation.

First and foremost, I would like to thank my supervisor, academy professor Paavo Alku, for his guidance and support during this work. Paavo’s door has always been open for questions, concerns, and discussion. And never has he ever ordered you to do something...even though he is almost always right.

Paavo also connected me with professor Tom Bäckström who graciously allowed me to work with him for six months at the International Audio Laboratories Erlangen in Germany. While this research visit didn’t directly advance my thesis, it did enhance my capabilities as a researcher. For that, I’d like express my gratitude to Tom.

I would also like to thank the co-authors of the articles in this thesis. Especially Hannu Pulakka, who can help you solve any problem you might have; Ulpu Remes, who is so excited about machine learning techniques that some of that excitement and knowledge might just latch on to the listener; and Santeri Yrttiaho and Marko Takanen, who have diligently done the statistics for several of the papers (and taught me a lot in the process). I have also had the privilege of collaborating and traveling with many other colleagues. In particular, the (scientific) level of several conference trips has been clearly improved by the company.

My gratitude also goes to the pre-examiners of this thesis, professor Martin Cooke and professor Tim Fingscheidt, for providing very insight-
ful comments and corrections that made me look at the text from a completely different angle. I would also like to thank professor Gernot Kubin for agreeing to be my opponent.

During the time I have spent at the university, I have conducted around 300 individual listening tests. This has required a lot of participants, and at some point, I felt like I was asking everyone I knew or just saw around. I would like to thank all of the kind people who answered yes. The finalization of this thesis would not have been possible, if it was not for you.

Finally, I would like to thank my family and friends who have not always fully understood what is it exactly that I do, but have always helped me in any way they could. And kindly listened (without falling much asleep) as I have again and again explained about some frustrating problem in the fascinating world of speech processing.

Helsinki, October 16, 2017,

Emma Jokinen
Contents

Preface 7

Contents 9

List of publications 11

Author’s contribution 13

List of abbreviations 14

List of figures 17

1. Introduction 19

2. Background 23
   2.1 Speech production ........................................... 23
   2.1.1 Lombard speech ............................................. 24
   2.1.2 Clear speech .................................................. 26
   2.2 Speech perception and intelligibility ..................... 26
   2.2.1 Objective models of speech intelligibility ............. 28
   2.3 Speech transmission in mobile telephony ............... 29
   2.3.1 Simulation of a telephone connection .................. 32

3. Techniques for speech intelligibility enhancement 33
   3.1 Deterministic metric-based techniques ....................... 35
   3.2 Deterministic non-metric-based techniques ................. 37
   3.3 Statistical techniques ....................................... 39

4. Evaluation of intelligibility enhancement techniques 41
   4.1 Subjective evaluation techniques ............................ 41
   4.1.1 Intelligibility tests ....................................... 42
   4.1.2 Quality tests ............................................... 45
4.2 Objective evaluation techniques ........................................... 46

5. Summary of publications ................................................... 49

6. Conclusions ......................................................................... 59

References .............................................................................. 63

Publications ........................................................................... 79
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.


Author’s contribution

Publication I: “Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech”

The author was mainly responsible for developing the algorithm and for evaluating it with the exception of the statistical analysis. The author wrote the first draft of the manuscript except for the parts concerning the statistical analysis of the results (in Sections III-D, IV-A, and IV-B).

Publication II: “An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech”

The author was mainly responsible for developing the algorithm and for evaluating it with the exception of the statistical analysis. The author wrote the first draft of the manuscript except for the parts concerning the statistical analysis of the results (in Section 4).

Publication III: “Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech”

The author was mainly responsible for developing the algorithms and for evaluating them except for the training and objective evaluation of the GMMs and the statistical analysis. The author wrote the first draft of the manuscript except for some parts related to the aforementioned topics (in Sections 3 and 5.1).
Publication IV: “Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network”

The author was mainly responsible for developing the algorithm and for evaluating it. The author wrote the first draft of the manuscript.

Publication V: “Intelligibility enhancement of telephone speech using Gaussian process regression for normal-to-Lombard spectral tilt conversion”

The author was mainly responsible for developing the algorithm and for evaluating it. Most of the development for the GPR models was done by the second author. The author wrote the first draft of the manuscript except for parts in Section V-A.

Publication VI: “Phase modification for increasing the loudness of telephone speech in near-end noise conditions – evaluation of two methods”

The author was mainly responsible for developing the algorithms and for evaluating them. The author wrote the first draft of the manuscript.

Publication VII: “Intelligibility enhancement at the receiving end of the speech transmission system - effects of far-end noise reduction”

The author was mainly responsible for implementing the evaluation setup and for conducting the evaluations. The author wrote the first draft of the manuscript.
List of abbreviations

2LP  two-stage linear prediction
ACELP algebraic code excited linear prediction
ACR  absolute category rating
AI   articulation index
AMR  adaptive multi-rate
AMR-WB adaptive multi-rate wideband
CCR  comparison category rating
CMOS comparison mean opinion score
CVC  consonant-vowel-consonant
dnn  deep neural network
DRC  dynamic range compression
DTW  dynamic time warping
ERB  equivalent rectangular bandwidth
eSII  extended speech intelligibility index
GIF  glottal inverse filtering
GMM  Gaussian mixture model
GP   glimpse proportion
GPR  Gaussian process regression
kbps kilobits per second
logSD logarithmic spectral distortion
LP   linear prediction
LSF  line spectral frequency
LTAS long-term average spectrum
MOS  mean opinion score
MSIN mobile station input (filter)
List of abbreviations

PDF probability density function
PESQ perceptual evaluation of speech quality
QCP quasi closed phase (glottal inverse filtering)
RMS root mean square
SII speech intelligibility index
SNR signal-to-noise ratio
SPIN speech in noise
SRT speech reception threshold
STI speech transmission index
STOI short-time objective intelligibility
SUS semantically unpredictable sentence
SWLP stabilised weighted linear prediction
VCV vowel-consonant-vowel
WCR word-correct rate
WER word-error rate
List of figures

2.1 Pre- and post-processing of speech signals in a mobile device 31

5.1 The mean word-error rate scores in the narrowband condition in Publication I. ........................................ 50

5.2 Spectral tilt estimation using the two-stage linear prediction approach. ............................... 51

5.3 Flowchart of the normal-to-Lombard spectral tilt conversion algorithm. ............................. 52

5.4 Spectral tilt estimates obtained with different estimation techniques for narrowband speech. ........................ 53

5.5 The mean ratings on pressedness in Publication V. .......... 54

5.6 Synthetic speech waveform processed with the phase-modification algorithms ............................ 55

5.7 The simulated transmission system from Publication VII. . 56
1. Introduction

The ability to use mobile phones in any situation can often lead to a telephone conversation taking place in a noisy environment, such as a car or a restaurant. This often makes it hard for the listener to understand what the speaker is saying because the background noise is disruptive. In a natural, face-to-face communication situation in a noisy environment, the speaker would modify his or her speaking style according to the environmental conditions to ensure the fluency of the communication. However, with the widespread use of mobile communication devices, the speaker and listener are increasingly in completely different locations, and the speaker does not necessarily have any knowledge of the surroundings of the listener. In the absence of listener feedback, the speaker is unable to naturally adapt his or her speaking style in a way that the communication situation would require. Therefore, to ensure the fluency of speech communication in all situations, intelligibility enhancement is required in the mobile device.

The obvious solution for enhancing the intelligibility of speech is to simply increase the level of the speech signal, that is to say, the output volume of the device. While this approach works to some extent, both the characteristics of the loudspeakers as well as the unpleasant and painful sensations experienced by the listeners as a result of very loud sounds limit its applicability in high noise levels. Therefore, an alternative means of intelligibility enhancement is needed, where the time or frequency characteristics of the speech signal are modified while constraining the output energy or loudness to a comfortable level.

The aim of this thesis is to develop speech processing techniques that can be used for the intelligibility enhancement of telephone speech in near-end noise conditions. Near-end noise refers to noise in the listener’s environment as opposed to far-end noise, which is in the speaker’s en-
environment. The introduced algorithms are intended to be used in the receiving device of a communication channel as a post-processing stage right before the speech signal is played out to the listener. For this purpose, modeling of natural speech production and perception mechanisms with machine learning (for example) is used to modify speech signals to make them more understandable. Generally in speech technology, intelligibility enhancement algorithms can be used offline, which means that long segments of speech (such as entire sentences) can be modified as a single unit with a global energy constraint. This scenario is possible, for instance, in the intelligibility enhancement of synthetic speech, where the transfer of energy from one part of a synthesized sentence to other parts of the same sentence is allowed. However, the target application in this thesis, real-time speech communication, places restrictions on the delay caused by the processing and prevents excessive buffering of the signal. As a result, all of the techniques proposed in this thesis use frame-based processing in short, 10- to 20-ms frames with a local energy or amplitude constraint.

In this thesis, additional emphasis is placed on using aspects of human speech production as a base for algorithm development, in contrast to approaches proposed elsewhere, which to a large extent take advantage of characteristics of human speech perception. An important additional consideration is the evaluation of the proposed intelligibility enhancement algorithms. In order to understand how the developed post-processing methods affect speech intelligibility (and other attributes), subjective tests and objective metrics are used throughout the thesis.

In the first part of the thesis, human speech production and perception as well as different methods of subjective and objective evaluation are briefly described. Furthermore, an overview of telephone speech and related speech processing algorithms as well as previously proposed intelligibility enhancement techniques is given. The second part of this thesis contains the most significant publications from the author related to the topic of intelligibility enhancement. In Publications I, II, III, V, and VI, algorithms for intelligibility enhancement are introduced and shown to improve intelligibility in subjective evaluations. While in Publication I, a straightforward high-pass filtering approach is used, the algorithms proposed in Publications II, III, V, and VI are based on modeling human speech production. Publication IV belongs to the same category, and it introduces a feature extraction method for the estimation of the spectral tilt.
from telephone speech. The proposed algorithm is then applied in intelligibility enhancement in Publication V. A commonly adopted assumption in intelligibility enhancement, also used in Publications I, II, III, V, and VI, is that the input signal to the intelligibility enhancement algorithm is free of noise. In the case of telecommunication systems, this would mean that the speaker is in a quiet environment, which is not necessarily true in all situations. Therefore, in Publication VII, the impact of noise in the input signal on the performance of intelligibility enhancement algorithms is investigated.
2. Background

The purpose of this chapter is to lay out the main factors contributing to the intelligibility enhancement techniques and the corresponding evaluation methods discussed in later chapters. While the topics discussed in this chapter include large study areas, such as human speech production and perception as well speech transmission and environmental noise, they are only touched upon briefly in the narrow context of intelligibility enhancement.

2.1 Speech production

Human speech production is based on modifying the airflow originating from the lungs using the larynx, the vocal tract, and radiation at the lips and the nostrils to represent different speech sounds [181]. The function of the vocal folds, situated in the larynx, is to modulate and control the airflow originating from the lungs before it is passed to the vocal tract. In the case of voiced speech sounds, the vocal folds vibrate. The resulting glottal volume velocity waveform, which serves as the excitation to the vocal tract, is quasiperiodic. That is to say, it is almost periodic with $T_0 = 1/F_0$, where $F_0$ is the fundamental frequency [117]. The vocal tract acts as an acoustic filter on the glottal flow creating resonances, which are known as formants. The location of these resonances in frequency depends on the overall size of the vocal tract as well as its shape, which can be controlled by the movement of the articulators, such as the tongue and the jaw [181]. In addition to voiced sounds, speech sounds can also be unvoiced. The difference between unvoiced and voiced sounds is in the excitation: unlike for voiced sounds, such as the vowel /o/, the excitation for unvoiced sounds, such as the fricative /θ/, is generated by a turbulent noise in the absence of vocal fold vibration. In addition, the third category
of speech sounds is plosives, the production of which is characterized by a full constriction in some part of the vocal tract [33].

For speech technology applications, the speech production process can be approximated using a linear model, known as the source-filter model. This model depicts the vocal tract as a linear filter that is acting upon an excitation signal that can be voiced, unvoiced, or mixed [33]. Additionally, lip radiation can be approximated separately as a first-order differentiator or the effects of the radiation can be included in the excitation. Although the true excitation originating from the physical voice source has a varying spectral tilt, the excitation in most source-filter models is often considered to be spectrally flat. As a further simplification, the vocal tract filter is typically parametrized using an all-pole model, obtained by using (for instance) conventional linear prediction (LP), which means that zeros in the transfer function of the physical vocal tract are ignored.

The mode of phonation results from the position and the tension of the vocal folds as well as the amount of air flow [117]. While most applications in speech processing technology consider only the normal phonation type, human speakers are capable of considerably changing their speaking style. This includes different types of phonations, such as breathy and creaky phonation, as well as different speaking styles used in the presence of various communication barriers. The way human speakers naturally modify their speaking style in order to make themselves more easily understood is especially interesting from the point-of-view of intelligibility enhancement. Lombard speech [99], which is produced when talking in noisy conditions, and clear speech, which is employed for instance when talking to a hearing-impaired listener [19], are discussed in more detail in the following sections.

2.1.1 Lombard speech

The Lombard effect is observed when speakers are talking in noisy conditions and they try to make their speech more understandable. Although the Lombard effect is stronger in the presence of communicative interaction [11, 44, 77, 184], where the speaker is communicating with another person and receiving feedback from them, it is also observed when there is no communication involved, for instance, when the speaker is reading words aloud in a recording scenario with noise played to their headphones. The purpose of the speech modifications in this case is to aid in the self-monitoring of the vocal output to detect speech errors [44,103,156]. While
speakers can make voluntary changes based on this auditory feedback, ignoring it completely is not an easy task [153]. However, it has been shown that by increasing the level of side tone—the speaker’s own voice fed back to the ears—the intensity of the Lombard effect can be decreased.

Several types of modifications are observed in Lombard speech compared to normal speech. For instance, a decrease in spectral tilt [56, 101, 159, 174], durational changes [35, 56, 159, 174], a rise in $F_0$ [101, 159], changes in formant frequencies [159], and an increase in vocal intensity [35, 126, 159]. The strength and type of modifications have also been found to be different for male and female speakers [76], as well as for individual speakers in general. Also the type of distortion, that is to say, the characteristics of the background noise, has an impact on the overall strength as well as the type of changes observed in Lombard speech [44, 56].

As the level of the background noise in the speaker’s ears rises, the produced speech becomes more intelligible as the vocal effort is increased, unless the level of the noise becomes so high that it disturbs the auditory feedback [35]. With very high levels of vocal effort, such as in shouting, the intelligibility decreases [126,129,141]. Interestingly, Lombard speech has been shown to provide an intelligibility benefit for human listeners even with the intensity difference compensated for [35, 159]. This suggests that the intelligibility benefit obtained by Lombard speech does not result from an increase in energy alone, but rather it is a combination of several modifications, especially the decrease in spectral tilt and related spectral changes [27, 28, 101]. Some of the changes observed in Lombard speech, however, do not contribute to intelligibility, and they might be the by-product of the other modifications. For instance, the increase in $F_0$ has been shown to produce little intelligibility benefit [101] and it might result from the increased lung pressure which is needed to raise the overall intensity [9, 48, 102]. Also the effect of durational changes observed in Lombard speech were found to produce little intelligibility benefit [28]. In the case of automatic systems, such as speech or speaker recognizers, Lombard speech typically reduces the recognition performance due to the differences from normal speech that these systems are normally trained with [24, 55, 56, 76, 136].
2.1.2 Clear speech

Speakers use the clear speaking style when they are communicating with hearing-impaired listeners or when they are asked to speak as clearly as possible [119, 124, 125]. Speakers have also been shown to clarify their speech in response to a communication barrier experienced by the listener only [57]. The changes observed in clear speech contain both signal enhancements, intended to increase the clarity of the signal, and code enhancements, intended to increase the acoustic distance between conflicting sounds [19]. The modifications include decreased speaking rate, longer and more frequent pauses, increased speech intensity, increased $F_0$ range, and increased vowel space [125, 154, 159]. Similar changes have been observed in deliberately stressed or emphasized speech [30, 96].

Clear speech has been shown to provide an intelligibility benefit for both normal and hearing-impaired listeners, compared to normal speech [119], but the obtained advantage varies depending on whether the listeners are native or non-native [19], on the age of the listeners [41], and on the speaker [19]. While one of the major contributors to the intelligibility advantage of clear speech is the slower speaking rate [89], clear speech produced at normal speaking rates is still more intelligible than normal speech [89]. Changes, such as increased energy in the 1000 Hz–3000 Hz range and increased modulation depth of low frequency modulations of the intensity envelope, have been suggested to contribute to the intelligibility benefit [89, 90].

2.2 Speech perception and intelligibility

Human hearing works by transforming the sound pressure wave arriving into the ear, first into mechanical motion and then into electrical impulses in the auditory nerve, which are then transmitted to the brain [117]. The audible frequencies range approximately from 20 Hz to 20 kHz [133] with increased sensitivity in the frequencies between 3 kHz and 4 kHz resulting from the resonance of the outer ear canal present in this range [133]. The frequency separation of sounds takes place in the cochlea, where the vibrations of the basilar membrane are transformed to neural impulses. Each point along the basilar membrane has a characteristic frequency with which it resonates. Two types of hair cells are situated along the basilar membrane: the inner hair cells that code the mechanical re-
responses of the basilar membrane to the auditory nerve [109, 127] and the outer hair cells that can actively amplify the mechanical motion of the basilar membrane. [133]. The amplification done by the outer hair cells is strongest at low sound intensities, and results in increased frequency selectivity [109]. Due to the active amplification of the outer hair cells the system becomes non-linear with respect to frequency, time, and level [133]. The inner hair cells each have a different best frequency they respond to, but they also respond strongly to frequencies near the best frequency. This frequency region is referred to as the critical band or the auditory filter [133]. The critical bands can be described using the Bark scale or the equivalent rectangular bandwidth (ERB) scale [117].

The masking effect in frequency, referred to as simultaneous masking, where the presence of one sound interferes with the perception of another, concurrent sound, is related to the critical bands. Generally, in masking, the threshold of hearing for one sound is raised due to the presence of another sound. Specifically, in simultaneous masking, a lower-frequency sound is generally masking a higher-frequency sounds when they appear in the same critical band [117]. Also, the loudness perception of a sound can be approximated by combining the contributions of each of the critical bands and compressing the output [117].

In addition to simultaneous masking, non-simultaneous masking, where the two sounds are not concurrent in time, is possible. This phenomenon is divided into forward masking, where the masker starts before the desired sound, and backward masking, where the order is reversed and the desired sound starts first. Notably, forward masking is much more important because backward masking seems to be observed only with inexperienced listeners and only low intensity sounds are masked [133]. The non-simultaneous masking phenomenon is probably related to the temporal integration of energy in the auditory system [117]. For instance, backward masking, which in a naive interpretation might suggest that the auditory system violates causality, appears more reasonable when one treats sound perception as non-instantaneous, that is, requiring a buildup over time [190]. The masking effects can be utilized in the development of speech and audio processing algorithms to improve the quality and to reduce the computational complexity [133,181].

While the perception of speech is governed by the same basic principles as any other sound, listeners tend to use specific speech cues to aid in the recognition of individual speech sounds. Such cues include, for instance,
spectral transitions [43, 157], consonant regions in general [58], and the location and movement of the formants and vowel durations [41]. The correct understanding of larger linguistic units, such as words or sentences, is impacted by cognitive processing in addition to the perception of the individual sounds. For instance, the intelligibility of individual words increases as the size of the used vocabulary decreases [62, 128, 141], when the frequency of the word increases [60, 130], or when the length of the word increases [62]. The mathematical models used to estimate speech intelligibility introduced next are, however, mostly focused on low-level effects that affect the perception of individual sounds. Modelling the higher level processing accurately is a much more demanding task.

2.2.1 Objective models of speech intelligibility

Objective models of speech intelligibility can be used to predict how much a listener would understand of a corrupted speech sample by approximating or modelling the processing that occurs in speech perception. One of the earliest intelligibility measures is the articulation index (AI) [42, 91], which evolved into the speech intelligibility index (SII) [12]—one of the most widely used metrics today. The SII uses the long-term averages of the speech and noise spectra to compute a band-wise signal-to-noise ratio (SNR), where psychoacoustic effects such as masking and auditory thresholds are taken into account. These band-wise SNRs are then combined in a weighted average, where the weighting—the band-importance function—typically varies according to the speech material used. The weighted average, referred to as the SII, is an estimate of how much speech information is available to the listener.

While the SII metric provides good intelligibility estimates for stationary noise types, the prediction fails with fluctuating noise types [140]. Since the computation is based on the long-term average spectra of both the speech and noise, temporal variations—such as present in fluctuating noise—are not taken into account in the processing. As a solution, approaches such as the extended speech intelligibility index (eSII) [140] have been proposed. Instead of using the long-term averages of the speech and noise signals, in the eSII, an instantaneous SII value is computed in short time frames and then all of the instantaneous intelligibility values are averaged over time. This allows the model to also include temporal changes in the predictions. The glimpse proportion (GP) measure [29] also estimates the intelligibility similarly based on the proportion of spectro-
temporal regions, where the speech signal is more energetic than the noise signal. The excitation pattern used as input for the glimpse detection is derived from the envelope of the basilar membrane’s response to sound. Glimpse-based estimation has also been extended to a binaural model in [171].

While the previously presented models are intended for linear distortions, such as additive noise, the speech transmission index (STI) [155] can also be used for reverberation and other non-linear distortions observed in communication channels, such as clipping. For the estimation of STI, a probe signal is generated by modulating a noise signal on several frequency bands, and then this probe signal is applied on the communication channel. At the output, the changes in modulation depth are measured and used to determine the final STI metric.

Although the STI- and AI-based models provide accurate intelligibility estimates for a wide variety of conditions, they are not well-suited for all applications, for instance, for the evaluation of time-frequency weighted noisy speech [165]. Short-time objective intelligibility (STOI), which is based on correlating the temporal envelopes of the clean and distorted speech signals in overlapping time windows, has been designed for predicting intelligibility of noisy speech upon which time-varying gains have been applied [165]. Measures intended more for generic quality evaluation of distortions, such as the spectro-temporal measure proposed by Taal et al. [162] or the perceptual model proposed by Dau [32], have also been applied as objective intelligibility measures.

In general, objective intelligibility measures work well with unprocessed, natural speech both with stationary and fluctuating noise types, but their performance typically reduces for modified or synthetic speech [172, 178]. Techniques, such as eSII and GP, that are based on estimating how much of speech is audible tend to give overly optimistic predictions for fluctuating maskers and too low predictions for stationary maskers. On the other hand, techniques based on measuring distortion compared to a reference signal, such as STOI, show opposite behavior with respect to masker type with synthetic and modified speech [172].

### 2.3 Speech transmission in mobile telephony

For wireless speech transmission to be possible, the analog speech signal needs to be first quantized and then compressed for transmission.
The process of *encoding* takes place on the transmitting side of the communication channel, whereas on the receiving side the compressed signal is *decoded*. Furthermore, the speech signal is usually band-limited for transmission to reduce the number of bits required. The classical telephone band is referred to as *narrowband* and ranges from 300 Hz to 3.4 kHz. It covers the most important frequencies for speech perception, which means that narrowband speech is almost completely intelligible [181]. However, the perceptual quality is significantly improved if the bandwidth is increased to *wideband*, which covers frequencies from 50 Hz to 7.5 kHz [110]. Even wider bandwidths include *super wideband* (50 Hz–14 kHz) and *fullband* (20 Hz–20 kHz). However, the increase in perceptual speech quality starts decreasing after bandwidth is increased beyond wideband [137], and this is often considered to be a sufficient frequency range for speech applications [167].

The speech signals are typically compressed for transmission using algebraic code excited linear prediction (ACELP)-based codecs. The main idea is to transmit the quantized LP spectral envelope along with an entry from the residual codebook that has been selected using a perceptually weighted error criterion. The most commonly used speech codecs include adaptive multi-rate (AMR) and adaptive multi-rate wideband (AMR-WB), which have also been used in the publications in this thesis.

In addition to the speech coding and decoding, the transmitting and receiving devices can do multiple pre- and post-processing operations to the speech signal to enhance its quality and intelligibility, as shown in Figure 2.1. The possible *pre-processing* steps—taking place in the transmitting device before the speech is encoded for transmission—include beamforming, acoustic echo cancellation, and noise reduction. Beamforming (e.g., [10, 100, 186]) is intended to enhance the speech signal by utilizing at least two microphones and taking into account the spatial distribution of the desired signal and the interfering signals [181]. Acoustic echo cancellation (e.g., [16, 36, 53]), on the other hand, tries to suppress the leakage from the loudspeaker to the microphone with an adaptive filter to prevent the speaker at the other end of the communication channel from hearing a delayed version of his or her own voice [181]. The third pre-processing stage includes noise reduction (e.g., [18, 37, 95, 105]), which is used to suppress the *far-end environmental noise* present in the desired speech signal.

The *post-processing*—done at the receiving device after the speech sig-
nal has been received from the communication channel and decoded—can include bandwidth extension as well as other quality and intelligibility enhancements. Bandwidth extension (e.g., [14, 69, 131]) is applied to expand the speech signal from a limited frequency band (typically narrowband) to a wider band (most often wideband). The quality of the received speech signal can also be enhanced using techniques referred to as post-filtering, which traditionally means that an adaptive filter enhancing the spectral peaks and suppressing the valleys is utilized to reduce the effects of quantization noise (e.g., [23, 49, 113]). The focus of this thesis, however, is on intelligibility-enhancing modifications applied on the received speech signal as a post-processing step. The purpose of the post-processing in this case is to modify the speech signal such that it is easier understand when interfered by near-end noise that is in the environment of the receiver. In this context, bandwidth extension could also be used to improve speech intelligibility [15, 92]. However, the focus of this thesis is solely on techniques where the processing is limited to the original signal bandwidth (either narrowband or wideband) encoded by the mobile phone sending device. Bandwidth extension is therefore left outside the scope of this thesis.

All of the pre- and post-processing present in the speech communication chain are restricted by tight delay requirements. In a natural conversation, the speaker expects a response from the other participant within a certain time and notices delays that are outside of this time window [82]. In real-time speech transmission, a mouth-to-ear delay—meaning the time
Background

from when the speaker has uttered a sentence to when a response is heard—of over 150 ms has been shown to lower user satisfaction [67].

2.3.1 Simulation of a telephone connection

All of the publications in this thesis are focused on post-processing telephone speech in the receiving device of a communication system. For the speech data to be realistic, a simulated telephone connection is used before the data is passed to the post-processing algorithms. As guidelines for the simulation, specifications detailing how data should be prepared for subjective evaluations of telephone connections (e.g., [7]), have been used.

The input speech, which is typically a high-quality recording with a 48-kHz sampling rate, is first downsampled to 16 kHz and filtered with the mobile station input (MSIN) [68] filter. The MSIN filter is used to simulate the characteristics of the mobile station input. Depending on whether narrowband or wideband speech is used, the MSIN-filtered speech is either downsampled to 8 kHz or maintained at the 16-kHz sampling rate. For narrowband, the speech signal is encoded and decoded using the AMR codec [2, 3] with a 12.2-kilobits per second (kbps) bit rate, whereas for wideband, the AMR-WB codec [1, 4] with a 12.65-kbps bit rate is used. The output of the decoder is equalized with SV56 [64,68] before being used as input to the post-processing block. After the post-processing, noise is added to the samples to simulate the distortion caused by additive near-end noise. In Publication VII, far-end noise, added to the signal before the MSIN filtering, was also used.
3. Techniques for speech intelligibility enhancement

The ultimate goal of speech intelligibility enhancement is to make the processed speech signal more understandable. Usually this is achieved by modifying the speech signal in such a way that it becomes more robust to interference, which is most often near-end environmental noise. In a general intelligibility enhancement scenario, where speech is played out from a loudspeaker, for instance, in a train station, the speech signal is often distorted both by reverberation and environmental noise before it reaches the listener. To combat the negative effects of the interference on intelligibility, the speech signal is modified by a post-processing block\(^1\) to enhance the signal that is played out to the listener. Typically, the speech signal is considered to be free of far-end noise and other distortions, such as reverberation, meaning that the post-processing block always has a clean input. However, this assumption is not entirely realistic, and the effects of noisy input speech with or without noise suppression on the performance of intelligibility enhancement techniques have been studied, for example in [50] and in Publication VII.

A post-processing algorithm typically consists of an analysis stage of the incoming speech signal followed by a modification stage. The target of the modification is to increase the intelligibility of the speech signal, and it can be based on a statistical approach (for instance, a Gaussian mixture model [79, 108]), or on a deterministic approach (such as optimization of an intelligibility metric [146, 166]). The actual processing of the speech signal can be implemented in both approaches using, for instance, filtering in either the time or frequency domain. That is to say, the main difference between the statistical and deterministic approaches is in how the parameters used in the processing (for example, the coefficients of the

\(^1\)While many studies refer to this modification stage as a pre-processing stage before the playback, here it is referred to as a post-processing stage to keep consistency with the mobile communication scenario presented in Section 2.3.
Techniques for speech intelligibility enhancement

In this thesis, different post-processing techniques are classified into three categories. They are first divided according to whether a deterministic or statistical approach is used. Then the deterministic approaches are further categorized into metric- and non-metric-based techniques. Metric-based approaches are typically based on mathematical optimization of speech intelligibility metrics, whereas non-metric-based techniques include information about speech perception and production into the intelligibility enhancement algorithm without optimizing intelligibility metrics.

Another important part of an intelligibility enhancement algorithm is the normalization of the energy of the post-processed frame to the energy level of the incoming frame. This is referred to as the energy constraint. Commonly, the normalization is done in terms of energy (e.g., [164, 168]), but also signal amplitude can be used, as done in [75, 139] and in Publication VI. The normalization can be done as a final phase of the post-processing algorithm, but the constraint can also be included already in the modification itself, especially in the case of deterministic metric-based techniques. The choice of the time scale used for the normalization has a large impact on the behavior of the algorithm. If a short time scale is used, speech energy cannot be efficiently redistributed along the time axis, while a longer time scale enables this. The redistribution of speech energy over time can be useful, for instance, in the case of a non-stationary noise signal or for boosting low-energy signal sections. However, the longer time scale also increases the delay and thus restricts the usability of the algorithm in applications with tight delay requirements.

In the following, the approaches adopted in the three categories of post-processing techniques—deterministic metric- and non-metric-based approaches and statistical techniques—are described further. In addition to techniques intended for standalone speech intelligibility enhancement, some relevant algorithms proposed in the context of speech synthesis applications are also included. While the focus of this thesis (and therefore also the discussion here) is restricted to intelligibility enhancement techniques that consider only environmental noise, algorithms for scenarios that also include reverberation are considered, for instance, in [31, 59, 85, 120, 123, 147].
3.1 Deterministic metric-based techniques

Metric-based intelligibility enhancement algorithms take advantage of models of human hearing and speech intelligibility (such as the ones introduced in Section 2.2.1) to formulate an optimization problem. In addition to the low-level intelligibility models working on the acoustic signal waveform, metrics using a higher level of abstraction, such as the word or sound sequence, can also be used [83]. For instance, word-level transcriptions and a statistical model of speech are used in [121] to derive a frequency-band-specific weighting applied on the word level. Another approach proposed in [187] is based on paraphrasing, where the linguistic content of the message is adjusted to make it more understandable. The latter approach considers the highest level of abstraction, where the actual sounds are not considered relevant but only the intended message is important. Most of the metric-based algorithms, however, consider low-level models utilizing the acoustic signal waveform and use this to derive a spectro-temporal weighting, which can then be used to modify the speech signal.

The low-level objective metrics used include the SNR, the SII [12], the GP [29], the spectro-temporal distortion measure [161, 162], and perceptual loudness [111, 112]. For instance, bins in time or frequency with the SNR below a set threshold level are amplified in [143, 168, 169]. A similar approach, but with the output power constrained to the input power using a short time scale, is proposed in [142]. In practice, this constraint forces the algorithm to suppress some frequency bands while other bands become boosted, which results in a more efficient use of the allocated energy.

Although the predecessor of SII, the AI, was used already in [52] to derive an optimal linear post-filter for intelligibility enhancement, the SII was first proposed as a tool for determining intelligibility-maximizing frequency-band boosting in [144]. Following a numerical solution [145], a recursive closed-form solution using a linear approximation was proposed [146], which was later improved with a nonlinear approximation that is more accurate at low SNRs [166]. In [139, 148], the SII-dependent frequency weighting is combined with dynamic range compression (DRC), which was also controlled using SII estimates. The SII has also been used as an additional cost in unit selection synthesis to maximize the intelligibility of the synthesized speech [22].
Optimal reallocation of speech energy based on the GP is studied in [13, 168, 170]. While selective boosting of frequency bands is considered in [168, 170], temporal reallocation is proposed in [13], and insertion of pauses at word boundaries based on optimizing the GP was proposed in [169]. The GP has also been used in statistical speech synthesis to modify the Mel-cepstral coefficients generated by a statistical model trained on normal speech to increase the intelligibility of the resulting synthetic speech [179].

The spectro-temporal auditory model is considered in [163, 164], where redistribution of speech energy in both frequency and time domain is allowed. The distortion model used in the optimization also takes into account temporal effects which results in efficient enhancement of speech transients. However, adapting the algorithm to short delay requirements decreases the intelligibility benefit obtained from the temporal processing [164]. A model of psycho-acoustic masking is considered in [21] as a metric in the computation of optimal gains for the first three formants. Loudness models are applied in the optimization of speech loudness for monaural scenarios in [152] and for binaural listening in [150]. While most of the algorithms for intelligibility enhancement are intended for applications enabling binaural perception (for example, for train station or airplane announcements), this aspect is not commonly considered in the processing.

In addition to the commonly used objective intelligibility metrics and perceptual models introduced in Section 2.2.1, other objective metrics have also been used for deriving optimal transforms for intelligibility enhancement. In [84], a new model for speech communication is proposed, where both the message production and interpretation processes are considered noisy. The measure of intelligibility of the received message, which is obtained from the mutual information rate between the original message and the received message, resembles a scaled SII. By maximizing the information throughput with both production and interpretation noise included in the model, better performance is obtained than without including them. In [122], a distortion measure to characterize the deviation of the spectral dynamics of the noisy speech from those of natural speech is proposed and used for optimization.

In addition to the incoming speech signal, an estimate of the background noise is also required to be able to fully take advantage of the objective metrics described. The noise conditions can be assumed to be known (e.g., [163, 168]) or they can be estimated using a noise tracker as in [164].
Since the original undistorted speech signal is available to the algorithm, the estimation of noise statistics is a much easier task than in the case of noise suppression algorithms, where only a distorted version of the speech signal is available [163, 164].

3.2 Deterministic non-metric-based techniques

Whereas metric-based techniques often utilize objective metrics and mathematical models of human speech understanding to formulate optimization problems, the non-metric-based approaches use aspects of speech perception and production in the form of expert knowledge combined with analysis of suitable speech data. This typically means a database containing both normal speech\(^2\) and clear or Lombard speech. For instance, in [87], clear and normal speech data is analyzed, and based on the observed differences, a modification with different weights for different frequency bands is proposed. In general, individual parameters, controlling for instance the amount of boosting in certain types of speech sounds or the changes in duration for vowel sounds, can be obtained in several ways in non-metric-based techniques. One method is to analyze suitable speech data, such as Lombard or clear speech (e.g., [46, 87]). Another approach is to use an objective intelligibility metric (such as the AI, SII, or GP) to tune a parameter (e.g., [87, 138, 139]). As a final stage, independent of how a parameter value has been first obtained, fine tuning can be used, for instance, by using informal listening (e.g., [34, 189]).

The arguably most straightforward approach for intelligibility enhancement is to use some form of high-pass filtering. This type of approach is motivated by the transfer of speech energy from the low-frequency range, where most real-life noises typically have most of their energy, to higher frequencies, where the effect of the noise is reduced. Approaches for intelligibility enhancement based solely on high-pass filtering have been proposed, for instance, in [52, 54, 154]. The high-pass filtering has also been used in combination with a low-pass filter intended to suppress excessive boosting of the frequencies above 6 kHz [149] and with additional techniques intended for boosting low-energy speech sounds, such as clipping [175] or amplitude compression [115]. Several studies also apply only the enhancement of different low-energy speech parts that are eas-

\(^2\)Normal speech in this context means speech of a normal phonation type recorded without background noise.
Techniques for speech intelligibility enhancement

ily masked by noise to improve the intelligibility of the speech signal. For instance, the amplification of consonants [58, 154], unvoiced speech sounds [94], stop consonants [70], and transients [138, 173, 185] have been proposed.

Whereas high-pass filtering is based on reallocation of speech energy in the frequency domain, energy reallocation can also be done in the time domain. Especially techniques intended for the boosting of low-energy speech sounds typically utilize this type of energy reallocation. For instance, DRC amplifies low-amplitude sections, such as unvoiced sounds and transients, while suppressing high-amplitude sections of the speech signal, such as vowel sounds. When this operation is combined with energy normalization, typically done over a longer time period, some of the speech energy originally allocated to the high-amplitude speech parts is in a sense reallocated to low-amplitude speech parts. Constraining the energy to the level of the original speech signal within a short time window, however, does not allow this kind of energy reallocation in time. To avoid excessive delay that prevents real-time processing, a running estimate of the global root mean square (RMS) energy can be used for normalization in shorter time windows [176].

Other types of non-metric-based techniques motivated by how speech is perceived and how environmental noise masks the signal include time-scale modifications [25, 71, 114] and enhancement of speech modulations that are important for intelligibility [88, 118, 182]. In addition to methods focused on speech perception, non-metric-based approaches motivated by modifications observed in different speaking styles, such as Lombard and clear speech, have also been proposed. Techniques such as formant sharpening [39, 63, 188], vowel-space expansion [45, 47], spectral envelope transformation [38, 46, 178, 188], time-scale modification [39, 63, 86], and $F_0$ modification [39, 63, 178] have been used. The modifications mimicking human speech production can also be used in combination with more straightforward approaches, such as DRC (e.g., [45, 86, 188]), or they can be enhanced to produce a much stronger effect than observed in natural speech to obtain further gains in intelligibility. Several Lombard-speech motivated modifications are also applied in a non-metric-based approach in Publication II, while in Publication VI a phase-modification resulting in sharpened excitation is discussed.
3.3 Statistical techniques

In contrast to the deterministic approaches introduced this far, statistical techniques take advantage of the distributions of speech parameters or features. For instance, a Gaussian mixture model (GMM) can be used to approximate the joint probability density function (PDF) of normal and clear speech features [108], and this statistical model can be used to map normal speech features to clear speech features. The features can be for instance the all-pole model of the spectral envelope parametrized as line spectral frequencies (LSFs) [79, 108] or formant frequencies [80, 108]. In addition to a suitable feature representation, the statistical approaches require training using appropriate speech data. This allows the model to learn what changes in the input data affect the output data and in what way. For training a GMM to model the dependency between normal and clear or Lombard speech, corresponding feature vectors have to be obtained in both speaking styles using, for example, dynamic time warping (DTW) [79] or manual labeling of speech data [108].

The approach is similar to voice conversion [158] where the features of one speaker are mapped to those of another speaker, except in this case the target is to maintain the speaker identity and to only change aspects of the speaking style to make the speech more intelligible. This kind of style conversion can also be applied to synthetic speech [93, 160], while another option is to adapt the synthesis parameters directly to the target speaking style [135, 180]. Publication III discusses a GMM-based transformation technique for Lombard speech, while Publication IV and Publication V propose, respectively, a feature estimation technique and Gaussian process regression (GPR) mapping for the Lombard modelling problem.
4. Evaluation of intelligibility enhancement techniques

Although an algorithm designed for speech intelligibility improvement might work well in theory, the final target audience of intelligibility enhancement is always the human listener. Therefore, it is important to use actual listeners to evaluate the performance of a speech processing algorithm. While subjective tests are arguably the best method of evaluating speech processing algorithms, they are often considered costly and time consuming to conduct. Furthermore, the results can show large variations depending on aspects such as the native language, the mental focus, and individual preferences of the listeners. Therefore, objective metrics can be used either as preliminary measures to adjust the performance of an algorithm, to complement the subjective tests, or as a standalone evaluation.

The purpose of this chapter is to describe the most commonly used subjective and objective evaluation techniques in intelligibility enhancement. Naturally, the focus is on methods for evaluating intelligibility, while some of the most important quality evaluation techniques are also described. While these two aspects can be considered related, that is to say that quality can be regarded as a generic umbrella term that also includes intelligibility [107,183], intelligibility can also be considered as a separate entity [98] and evaluated as such. This is also the view adopted in this thesis and all the publications in it.

4.1 Subjective evaluation techniques

Intelligibility is the most important attribute considering the topic of this thesis. While it certainly is important for speech communication in general, speech communication could also be evaluated in terms of complex characteristics, such as conversation effectiveness or ease of communica-
tion [107]. However, usually in subjective evaluations, listeners from the target group are asked to judge the performance of the proposed algorithm with respect to a single, distinct attribute that can be easily specified. Other, more quality-related attributes include for instance naturalness, loudness, and annoyance.

Subjective evaluations are most often conducted in a sound-insulated room with high-quality headphones. For instance, many standardized quality tests specify these type of listening conditions [65] with the purpose of controlling the test environment better and allowing the listener to focus only on the task at hand, thus making the result more reliable. Although most intelligibility tests are not governed by standards in the same way, the presence of noise disturbing the test can have an even more dramatic effect on the results. In most intelligibility tests, each test sample is repeated only once to avoid learning effects. If an unfortunate background noise disturbs the listener at the wrong moment, the resulting intelligibility score might be affected.

However, compared to real-life listening situations, headphone listening in an isolated listening space is far from realistic. For instance, in a normal communication situation with a mobile phone, different types of environmental noise and other parallel tasks, such as driving, might disturb the listener and increase his or her cognitive load. In this kind of situation, both the quality and intelligibility of speech might be perceived differently from an ideal test scenario. Possible approaches to increase the correspondence to real communication situations include the use of loudspeakers to playback the noise [132] or to playback both the noise and speech [134], or giving other concurrent tasks to the listener [106].

4.1.1 Intelligibility tests

As mentioned in Section 3.1, intelligibility can be examined on several levels of abstraction. Subjective tests are typically focused on either microscopic or macroscopic intelligibility, where the former considers individual phonemes and the latter considers words and sentences [141]. Also, the highest, message-level abstraction could be applied naturally for testing purposes, for instance, in some type of conversational or other type of tests involving continuous speech (e.g., [104]), but in practice it is not commonly used. While the macroscopic intelligibility corresponds more closely to natural communication than the use of individual phonemes, the effect of context especially affects the sentence level measurements [20, 81].

42
means that the listener can take advantage of the syntactic and semantic information as well as possible coarticulation effects within a sentence to correctly deduce some of the words, even if they did not hear them completely [20]. As mentioned in Section 2, the frequency of words as well as the size of the vocabulary also have an impact on the intelligibility scores. These factors should therefore be taken into account in the design of the test procedure and material.

Probably the most common test type for evaluating speech intelligibility enhancement algorithms is a straightforward word-error rate (WER) measurement where a corrupted sentence is played out and the listener is tasked with typing the sentence word-for-word. The level and type of noise corrupting the speech is typically kept constant throughout the entire test or at least for a block of sentences. If the background noise is changed, the listener is allowed to adjust to the change with a set of training examples with the new background noise. The measure of intelligibility in this type of test, referred to as WER, is simply computed as the percentage of unidentified words averaged over a set of sentences. The opposite measure, the word-correct rate (WCR), can be reported instead. This type of test procedure has been used in [45, 122, 170, 177] and in all of the publications in this thesis except Publication IV.

Another variety of intelligibility measurement is the speech reception threshold (SRT) test, where the procedure is similar to the WER test except that the noise level is adapted constantly throughout the test. The test samples are presented in blocks, and the level adaptation for each sample in the block is done on the basis of the percentage of correct words recognized from the previous sample. The purpose of the adaptation is to determine an SNR level where a certain percentage of speech (typically 50%) is understood, and the final measure reported from the experiment is the average SNR level computed from the samples at the end of the block. The benefit of this type of adaptive procedure is that floor and ceiling effects, where the samples are either fully intelligible or completely unintelligible, are avoided [116]. On the other hand, a single measurement point typically requires a block of around 15 samples, and thus the data collection might require more effort. Additionally, the adaptive procedure requires an automated determination of the WER score on the run, which can be sensitive to typing errors. The SRT test procedure has been used in [21, 78, 148].

The sentence material used in these two test types can be selected de-
Evaluation of intelligibility enhancement techniques

pending on whether or not the use of context is desirable. While natural sentences contain semantic and contextual cues that the listener can use, test material can also be constructed to minimize these effects. For instance, semantically unpredictable sentence (SUS) material is constructed by using fixed syntactic structures where words are inserted randomly from the specified groups to form new sentences [17]. This makes the resulting sentence unpredictable while still being syntactically correct. The so-called matrix test (e.g., [61]), where a test sentence is formed by selecting one word from each column of a matrix of words, also typically results in SUS material. This type of low-context sentence material has been used in evaluating intelligibility enhancement algorithms, for instance in [58, 139, 166].

In addition to using full sentences as test material, the recognition of smaller units, such as words or phonemes, can also be evaluated. These units can be presented within a carrier sentence or embedded in a word, but it is also possible to present the key word or phoneme without any carrier. However, for individual phonemes, important cues for recognition lie also in the transitions from the neighboring phonemes which are not present if a carrier is not used. An example of a word-recognition test that uses carrier sentences for embedding the keywords is the speech in noise (SPIN) test [81], where the type of context—either low or high probability—can also be controlled by the selection of the carrier sentences.

Most commonly used carriers for individual phonemes include for instance meaningful or nonsense words, both of which are typically designed to provide little context information to the listener. Nonsense words can be constructed using the consonant-vowel-consonant (CVC) or vowel-consonant-vowel (VCV) structure for instance, where the vowel or consonant of interest is surrounded on both sides. While these type of nonsense words do not give much context information about the center phoneme, meaningful words, on the other hand, most often contain cues that can aid the listener in guessing the correct word even if the entire word is not heard properly. To reduce the listeners’ ability to fill in the missing phonemes based on the neighboring ones, the set of test words can be designed to be confusable such that a pair of words differs only in one phoneme. For instance, in the rhyming tests (e.g., [40, 51]), the test material consists of groups of rhyming words that differ only in the initial or the final consonant. The task of the listener is then to identify the word.
that was played from the group of confusable words. Rhyming tests have been used for intelligibility evaluations in [26, 54, 173], while tests with nonsense words have been applied in [58, 70, 108].

While the intelligibility tests discussed so far are subjective in the sense that human listeners are required in conducting them, the test scores are objective to some extent. That is to say, the intelligibility scores are not based on the opinions of the listeners but on the content they have reported to have heard. While there are still listener-dependent, subjective effects present in the results, the personal preferences and biases of the listeners are mostly removed. However, intelligibility can also be evaluated in a purely subjective way by asking listeners to grade the intelligibility of samples on a scale or to choose which sample in a pair of samples is more intelligible. The main benefit in this kind of approach is that it typically takes less time, at least compared to the WER and SRT tests. Intelligibility tests based on a rating scale have been used in [86, 149], while Publication I applies a pair comparison test to measure intelligibility of post-processing algorithms.

4.1.2 Quality tests

Subjective tests designed to measure attributes such as naturalness, loudness, or the overall quality are based on the listeners’ opinions about the test sounds. That is to say, the test result strongly depends on the subjective views of the listeners and on their interpretation of what is meant by the attribute of interest. This variation can be somewhat reduced by good test design, where the task is as self-explanatory as possible and important aspects to pay attention to are reiterated to the listeners beforehand, but there is always room for interpretation. This often leads to a large number of listeners required for reliable results; for instance, in speech codec evaluations, approximately 30 listeners are typically used per test condition [5, 6].

The most common quality test types include the absolute category rating (ACR) and the comparison category rating (CCR) [65] tests. In the ACR test, the listeners are asked to rate a speech sample with respect to a specific attribute, which is typically overall quality. The scoring is done on a five-point scale ranging from bad to excellent, and the test result—mean opinion score (MOS)—is computed as an average over all the ratings. In the CCR test, the listeners are presented with two test samples, and they are asked to rate the second sample compared to the first one. The rating
Evaluation of intelligibility enhancement techniques

is done on a seven-point scale ranging from much worse to much better, resulting in the comparison mean opinion score (CMOS). The CCR test can also be modified to a pair comparison test where the preferred sample is indicated using either forced choice (one of the samples has to be selected as preferred) or with a “no preference” option provided. The ACR has been used in [23, 113] and the CCR in [54, 151, 152], while the pair comparison test has been employed in [164, 189] and in all publications in this thesis except for Publication IV.

4.2 Objective evaluation techniques

Even if subjective tests are not always conducted using entirely realistic communication scenarios, they are still the best method of estimating how the end users will perceive the speech processed by a new algorithm. On the other hand, subjective tests can be time- and money-consuming. Therefore, especially for a preliminary evaluation preceding a subjective test, objective tests provide an attractive, low-cost alternative.

Objective evaluations are conducted without the aid of human listeners. Instead, a metric that is somehow correlated with human intelligibility or quality perception is computed from the processed speech samples with or without the aid of the original speech sample. The correlation between listening test results and the scores provided by an objective metric are usually not perfect, but the results typically give indications of how the method would perform in a listening test. However, in some cases, objective metrics and subjective tests have shown contradictory results [148].

The most relevant metrics regarding the topic of this thesis, intelligibility enhancement, have already been introduced in Section 2.2.1 with the discussion on objective measures of intelligibility. While many of these metrics are used for algorithm development (see Section 3.1), they have also been widely employed for evaluation purposes. For instance, SII has been used in [139, 142, 166], eSII in [59, 86, 189], GP in [87, 170, 177], and STOI in [31, 164, 169] for evaluating intelligibility enhancement algorithms.

In addition to objective models of intelligibility, objective quality evaluations, such as perceptual evaluation of speech quality (PESQ) [66], can also be used. PESQ requires both the processed and the original speech signal and applies a perceptual and a cognitive model to derive an estimate of the listening quality in terms of MOS. This has been used for the
quality evaluations of intelligibility enhancement techniques in [164, 168, 182] and in Publication VI.
5. Summary of publications

This section summarizes the publications in this thesis.

Publication I: “Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech”

In Publication I, a post-processing technique intended for intelligibility enhancement in near-end noise conditions is introduced. The algorithm falls in the category of deterministic non-metric-based techniques. It uses an adaptive high-pass filter followed by adaptive gain control to transfer speech energy from low-frequency regions to higher frequencies. The post-filter consists of three cascaded filters where the first two filters are so-called formant filters and the third one is a tilt filter. The formant filters are used to modify the approximate frequency regions of the first and second formants such that the first formant is suppressed while the second formant is enhanced. The parameters of the filter structure controlling the amount of formant suppression and enhancement were derived using a listening experiment. The third cascaded filter, the tilt filter, is used to compensate for the possible changes in the spectral tilt introduced by the two formant filters. The proposed technique is also adaptive: in good background noise conditions, the amount of processing is reduced, and the effect of the processing is increased as the SNR decreases.

Subjective listening tests were arranged to evaluate the performance of the proposed technique in comparison to the reference method, the post-filter by Hall et al. [54], and unprocessed speech in both narrowband and wideband conditions. The tests consisted of a WER test and a preference test with questions regarding the subjective estimate of intelligibility, naturalness, and preference of the processing methods under evaluation. Results of the WER test in the narrowband condition are shown
in Figure 5.1. The test results indicated that in good SNR conditions, post-filtering does not cause distortions that would degrade intelligibility and is in some cases preferred over unprocessed telephone speech. While both techniques were able to increase intelligibility of both narrowband and wideband speech in difficult noise conditions, the proposed method received higher intelligibility and preference scores with narrowband speech in moderate-to-difficult noise conditions.

**Publication II: “An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech”**

In Publication II, another intelligibility enhancement approach designed for post-processing of telephone speech is presented. The algorithm belongs to the same category as the one presented in Publication I, but the approach aims to provide more natural-sounding modifications by incorporating aspects of human speech production, namely those observed in Lombard speech. The proposed algorithm consists of four sections: (1) spectral tilt compensation, (2) formant sharpening, (3) $F_0$- and noise-adaptive high-pass filtering, and (4) gain control. In the spectral tilt compensation, visualized in Figure 5.2, the spectral tilt of the input speech is first estimated using a two-stage linear prediction (2LP) technique and then removed using the resulting all-pole filter. The formant sharpening is achieved through iteratively updating the LSFs estimated from the tilt-compensated speech frame with the algorithm proposed in [97].
The high-pass filter used for processing is adapted from two filters based on an estimate of the $F_0$. These high-pass filters were derived from the difference between the long-term average spectra (LTAS) of normal and Lombard speech, one for male speakers and one for female speakers. The $F_0$-adapted high-pass filter is then windowed according to the SNR level to control the strength of the filtering.

The performance of the proposed intelligibility enhancement algorithm was evaluated in a WER test as well as in a pair comparison test with questions on suitability to noise conditions and preference in two noise conditions. The technique was compared to both unprocessed telephone speech and the post-filtering method by Hall et al. [54]. While both post-processing techniques improved intelligibility over unprocessed telephone speech, the proposed technique achieved lower WERs in difficult noise conditions. Furthermore, it was rated higher in terms of suitability to the noise environment in difficult background noise conditions.

Publication III: “Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech”

In the study reported in Publication III, a statistical post-processing technique for telephone speech in noisy near-end conditions is studied. The simplified modeling of Lombard speech production studied in Publication II is improved by using a statistical mapping for converting the spectral tilt of normal speech to that of Lombard speech. The proposed algorithm consists of several stages: spectral tilt estimation, trained mapping with a GMM, filtering, and gain control, as shown in Figure 5.3. For one of
the most important parts—the spectral tilt estimation—several different parametrization techniques (such as stabilised weighted linear prediction (SWLP) and 2LP used in Publication II) are implemented and evaluated in terms of how well the estimates match the LTAS computed from normal and Lombard speech. Finally, the spectral tilt estimation techniques are used to extract parallel feature vectors from read normal and Lombard speech. These features are then utilized for training GMM models. To ensure that the training data contains a representative Lombard effect, a data selection technique based on SII ranking is applied.

To study whether any of the spectral tilt estimation methods would be suitable for a post-processing application, they were evaluated in a WER test and a pair comparison test regarding overall quality with unprocessed telephone speech included as a baseline reference. The results of the intelligibility evaluations show that the post-processing algorithm with SWLP used for tilt estimation is capable of improving intelligibility of narrowband telephone speech.

**Publication IV: “Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network”**

A technique for estimating the spectral tilt of the glottal source directly from telephone speech using a deep neural network (DNN) is discussed in Publication IV. This type of spectral tilt estimate can be used in intelligibility enhancement, for instance, in the spectral tilt conversion from normal to Lombard speech, as presented in Publication III. While spectral tilt estimation can be achieved using all-pole techniques such as SWLP on the speech signal (as was done in Publication III), these approaches are not able to capture the true glottal excitation. Glottal inverse filtering (GIF), where the time-domain glottal waveform is estimated from speech, aims to solve a similar problem, but it cannot be directly used due to its sensitivity to distortions present in telephone speech. Therefore, the proposed method uses a DNN trained on reference spectral tilt models.
Figure 5.4. Spectral tilt estimates obtained with different estimation techniques for narrowband speech.

computed using quasi closed phase (QCP) GIF [8]. While the training inputs for the DNN are computed from telephone speech, the outputs are computed from high-quality speech that has only been downsampled to the correct sampling frequency.

The accuracy of the proposed tilt estimation technique was evaluated using a dataset of Finnish vowels with three different phonation types, where the reference spectral tilt was extracted using GIF. In addition to the proposed technique, two parametric all-pole estimation techniques (SWLP and 2LP) and a cepstral-based estimation (CF) were used in the evaluation. Figure 5.4 shows a comparison of the spectral tilt estimates obtained with these techniques and the reference spectral tilt obtained with GIF. The DNN-based technique achieved the lowest logarithmic spectral distortion (logSD) values in all phonation categories for both narrowband and wideband speech.

Publication V: “Intelligibility enhancement of telephone speech using Gaussian process regression for normal-to-Lombard spectral tilt conversion”

In Publication V, the statistical post-processing approach for intelligibility enhancement introduced in Publication III is studied further. The proposed conversion of spectral tilt from normal-to-Lombard speech is extended by introducing the use of nonparallel training data. Additionally,
the DNN-based spectral tilt estimation is applied and the GMM mapping is replaced with a GPR mapping that provided encouraging results previously [73]. The use of nonparallel, conversational Lombard data [74] poses new problems for locating matching input-output pairs for training the GPR mapping compared to read speech that is most often parallel. For this purpose, different matching techniques for finding corresponding normal and Lombard frames are studied. To further improve the quality of the training data—in other words, to ensure that the Lombard effect is well-represented and observable—different approaches for ranking the Lombard data are studied and used together with the matching techniques.

The performance of the proposed normal-to-Lombard spectral tilt conversion algorithm was evaluated in two separate subjective evaluations. The formant equalizing post-filter by Hall et al. [54] as well as the baseline tilt conversion technique from Publication III, were used as reference techniques in the evaluation. The first part of the evaluation consisted of a WER test in high-noise conditions and of a pair comparison test on overall listening preference in low-noise conditions. The second part of the subjective evaluation contained a smaller evaluation where subjects were asked to rate the pressedness of different speech samples compared to a reference sample in a multiple comparison test with a hidden reference without noise in the background. The results of the second test on pressedness are depicted in Figure 5.5. The results suggest that the modelling of the spectral tilt caused by the glottal excitation is perceptually different from high-pass filtering as measured by the pressedness test. Furthermore, the intelligibility improvement is still comparable to that of high-pass filtering.
Figure 5.6. Examples of synthetic speech waveform of the vowel /a/ processed with the phase-modification algorithms for both the narrowband and wideband conditions.

Publication VI: “Phase modification for increasing the loudness of telephone speech in near-end noise conditions – evaluation of two methods”

In Publication VI, the feasibility of phase-based modifications for intelligibility enhancement of telephone speech is studied. While magnitude spectrum or time-domain modifications are commonly exploited, phase-based modifications are not widely used. The key idea in the study is to increase the loudness of the speech signal by modifying the signal’s phase spectrum. Through the increased loudness, speech would stand out better from background noise, thereby resulting in enhanced intelligibility. Two separate phase-based modifications are studied: one is based on reducing the dynamic range of the signal combined with normalizing the amplitude to increase loudness, while the other method sharpens the high-amplitude peaks in the time-domain speech signal to make it sound clearer. While the former algorithm is based on an energy increase to achieve a rise in loudness, the latter approach is motivated by the changes in glottal excitation observed in naturally produced loud speech. After the phase modifications, frame-based amplitude normalization is used instead of the more frequently applied energy normalization. The effect of the two phase-modification algorithms on a synthetic speech signal is illustrated in Figure 5.6.

The proposed phase-modification methods were evaluated first in com-
parison to unprocessed speech using both objective and subjective metrics. The best performing algorithm for both narrowband and wideband was selected for a further comparison with DRC and unprocessed speech in subjective intelligibility tests in different noise conditions as well as subjective quality tests in clean conditions. The results suggest that both of the phase-modification methods are able to increase both subjective loudness and intelligibility compared to unprocessed speech and DRC in the tested bandwidth conditions. However, of the two techniques, only the peak increasing modification was rated higher in quality than DRC, while the peak reduction technique received lower ratings.

Publication VII: “Intelligibility enhancement at the receiving end of the speech transmission system — effects of far-end noise reduction”

In Publication VII, the effects of far-end noise and noise reduction on the performance of intelligibility enhancement techniques is studied. While the common assumption in near-end intelligibility enhancement is that the input signal is free from far-end noise, in practical situations this might not be true. As a result, the intelligibility enhancement algorithm might amplify noise or artifacts caused by noise reduction.

Three different post-processing algorithms were evaluated in subjective tests using the simulated transmission system depicted in Figure 5.7. The evaluation consisted of a WER test and a pair comparison test regarding quality with different far-end and near-end noise conditions. In addition to the post-processing algorithms, noise reduction was used in some of the far-end noise conditions to evaluate whether it affects the performance of the intelligibility enhancement. The results showed that far-end noise has a negative effect on the performance of the intelligibility enhancement techniques if noise reduction is not used. In general, noise reduction increased intelligibility in stationary far-end noise conditions. However,
the intelligibility gain obtained using post-processing compared to unprocessed speech was reduced. In non-stationary far-end noise conditions, the noise reduction did not provide any benefit in terms of intelligibility.
Summary of publications
6. Conclusions

Mobile phone customers can use their devices in almost any situation which means that environmental background noise often affects telephone conversations. The noise can disturb the communication and make it hard for the listener to understand what the speaker is saying. In a normal face-to-face discussion, the speaker would adapt his or her speaking style to overcome the communication barrier, but in a telephone conversation this requires listener feedback. While a simple increase in the output level of the device can be used to enhance the intelligibility of the speech signal, after a certain point this approach is no longer an option due to the limitations set by the loudspeakers. More advanced solutions, where the time and frequency-domain characteristics of the speech signal are modified to make it stand out better from the noise, are therefore required.

This thesis contributes to the development and evaluation of speech processing techniques that can be used to enhance the intelligibility of telephone speech. In Publication I, a robust, SNR-adaptive post-processing technique is introduced. This approach is based on transferring energy from the first formant to higher frequencies using a cascade of filters. Also in Publication II, a robust approach to enhancing the intelligibility of speech with post-filtering is proposed. However, this technique derives motivation from Lombard speech. The Lombard-based approach is then developed further in Publications III and V, where modification of normal spectral tilt to that of Lombard speech using machine learning techniques is studied. To capture the true spectral tilt of the glottal source accurately for the normal-to-Lombard conversion, a DNN-based spectral tilt estimation technique is proposed in Publication IV. The approach outperformed conventional techniques based on all-pole modelling in terms of logSD in all of the tested conditions.

Most often intelligibility enhancement techniques are based on modify-
Conclusions

ing either the short-time spectral magnitude of speech—as in Publications I, II, III, and V—or the time-domain signal directly. In Publication VI, however, the plausibility of phase-domain modifications for intelligibility enhancement is studied. One of the proposed phase modification techniques aims to increase speech loudness through an increase in speech energy: by first minimizing the maximum amplitude of the signal and then applying amplitude normalization, the energy is increased. The other phase modification technique is motivated by human speech production: by strengthening high-amplitude peaks in the time-domain speech signal caused by the closing phase of the glottal excitation, an effect taking place in the production of loud speech is mimicked.

All of the algorithms proposed in this thesis have been designed considering the tight restrictions on delay and computational complexity. While none of the algorithms has been implemented or tested in a real mobile phone implementation—at least to the author’s best knowledge—there are no fundamental obstacles preventing such an implementation. All of the proposed algorithms have also been evaluated thoroughly in subjective tests and have been shown to be beneficial in terms of intelligibility. For instance, compared to unprocessed speech, significant decreases in terms of WER ranging up to 40 percentage points were achieved.

While the algorithms proposed in Publications I and II provide efficient and robust methods of improving intelligibility, the main focus of this thesis is on incorporating aspects of human speech production to intelligibility enhancement. Even if the approaches presented in Publications III, V, and VI might not be as robust as simple high-pass filtering, the produced modifications should model more closely the acoustical characteristics of speech that is spoken by natural talkers in noisy environments. However, the Lombard effect, for instance, is an immensely complex phenomenon, and being able to model it fully in a real-time algorithm still requires more research effort. Despite being a hard task, the field of intelligibility enhancement would benefit in the future from not just being inspired by human speech modifications, but by actually trying to accurately model them. This direction of research might also deepen our understanding of speech production.

The last paper of the thesis, Publication VII, studies the effects of far-end noise and noise reduction on the intelligibility improvement provided by various algorithms. A typical, but unrealistic assumption made in the majority of the publications in the field of intelligibility enhancement, is
the lack of degradations in the input speech signal. In other words, the speaker is assumed to be in ideal conditions without any background noise or reverberation. While these types of distortions can be combated using different techniques as discussed in Section 2.3, the performance of the intelligibility enhancement might still be affected. The results in Publication VII show that while noise reduction increases intelligibility in stationary far-end noise conditions, the intelligibility gain obtained using post-processing compared to unprocessed speech is reduced.

An interesting venue of research pertinent to intelligibility enhancement is related to the subjective evaluation of intelligibility. Similarly to most of the publications in the field, the subjective evaluations conducted in Publications I, II, III, V, and VI have all taken place in ideal listening conditions with high-quality headphones. While this testing methodology can be used to observe differences between various algorithms, it is far from a real application scenario. While more realistic evaluation approaches exist as discussed in Section 4.1, they are not widely used. A further element of realism is the type of degradation used in the evaluation of intelligibility enhancement techniques. While near-end background noise is considered to be the most influential factor affecting intelligibility, modern communication systems also suffer from other degradations to the speech signal. For example, packet loss can lead to a reduction in intelligibility [72] and might require new approaches to tackle its effects.

Currently, most evaluation approaches, such as the WER tests used in Publications I, II, III, V, VI, and VII, require high levels of background noise in order to avoid ceiling effects in the intelligibility scores. However, even at lower noise levels, the cognitive load required to decipher the speech in noise is increased. While techniques for measuring the cognitive load experienced by test subjects exist, they typically require additional devices, such as eye-tracking cameras, and are not commonly applied in the evaluation of intelligibility enhancement algorithms. Therefore, a simple test design that would enable the estimation of the cognitive load and most importantly the comparison of different algorithms without any additional devices would be of use to the field of intelligibility enhancement.
References


References


[146] ——, “Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement,” in *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, 2010.


In mobile communications, the quality and intelligibility of speech can be degraded by quantization and coding. In addition, environmental noise might be present at the receiving or eavesdropping side of the communication channel. To combat the deterioration of intelligibility and quality of speech at the receiving or eavesdropping side, one can apply to the speech signal received from the communication channel. This thesis studies the impact of human listeners naturally take advantage of when talking in noisy situations. The performance of the techniques is demonstrated with subjective listening tests using simulated telephone speech and in general, the techniques show clear intelligibility improvement over unprocessed telephone speech.