TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Electrical Engineering

MARJA LÄHDEKORPI

PERCEPTUAL IRRELEVANCY REMOVAL IN NARROWBAND SPEECH CODING

Master of Science Thesis

Subject approved by the Department Council
February 12, 2003

Supervisors: Prof. Jukka Saarinen
M.Sc. Jani Nurminen
Prof. Ari Visa

Acknowledgements

The work for this thesis has been carried out at the Institute of Digital and Computer Systems, Tampere University of Technology (TUT). The work has been conducted in co-ordination with Speech and Audio Systems Laboratory, Nokia Research Center (NRC) as a part of the User-Oriented Information Technology (USIX) project. The work has been financed by the National Technology Agency (Tekes) and NRC. The examiners of this thesis were Professor Jukka Saarinen, M.Sc. Jani Nurminen and Professor Ari Visa.

First of all, I would like to express my gratitude to all the examiners of this thesis for the feedback and constructive suggestions. My special thanks go to M.Sc. Jani Nurminen for patiently concentrating on my work and for guiding and encouraging me throughout this project. I would also like to thank all the members of this speech coding project, both at TUT and at NRC, for the pleasant teamwork and continuous support. I am particularly grateful to M.Sc. Ulpu Sinervo for all the encouragement and technical assistance, not least with the listening test. In addition, I would like to thank Professor Jukka Saarinen for giving me the opportunity to work towards this degree at the Institute of Digital and Computer Systems.

Furthermore, I am very grateful to all the people who participated in the listening tests arranged within this work; thank you for your valuable contribution. I also appreciate the support given by my friends during my work on this thesis. In particular, I would like to thank my dearest friend Milla for all the consoling discussions that often helped to clear my mind of confusing thoughts.

I am also deeply indebted to my family for understanding and encouraging me during the time I have been working with this thesis. Especially, I want to express my sincere gratitude to Hannu for his endless patience and support. Finally, my ultimate special thanks go to my mother Anita: Thank you tremendously for all the encouragement and wise words of advice.

Tampere, April 30, 2003

Marja Lähdekorpi

Contents

TAMPERE UNIVERSITY OF TECHNOLOGY
Degree program in Electrical Engineering
Institute of Digital and Computer Systems
Lähdekorpi, Marja: Perceptual irrelevancy removal in narrowband speech coding
Master of Science thesis, 55 + 1 p.
Examiners: Professor Jukka Saarinen (TUT)
M.Sc. Jani Nurminen (NRC)
Professor Ari Visa (TUT)
Funding: Tekes
Nokia Research Center (NRC)
Department of Electrical Engineering
April 2003

The ulitisation of the properties of the human auditory system has a great potential as a means to enhance the efficiency in speech coding. Ideally, the bit rate can be noticeably reduced without causing significant speech quality deterioration, or better speech quality can be obtained within the available bit rate, by taking account of the operation of the hearing. This work concerns a generic method of improving the coding efficiency by modifying the speech signal based on perceptually appropriate criteria. The work has been done as a part of a larger research project that examines low bit rate speech coding.

The purpose of this thesis is to present a narrowband speech modification technique that makes the more efficient coding of the speech signal possible. The proposed method reduces the perceptual irrelevancy of the signal by removing the components that are imperceptible for human listeners. The objective is to avoid wasting the coding capacity on the irrelevant information. The operation of the technique builds upon a well-known auditory model that has originally been designed for audio signals. In this work, the model is adjusted for narrowband speech signals and used to calculate a masking threshold for each segment of speech. By comparing the threshold with the spectrum of the speech segment, the perceptually irrelevant masked frequency components are detected. These components are then removed by adaptive filtering.

It was deduced from the results of an informal listening test that the removal of the frequency components that were deemed masked by the model did not cause significant degradation of the speech quality. At the same time, objective measurements showed that the modified speech signal can most likely be coded more efficiently than the original signal. Another perceptual evaluation procedure was also conducted with two standardised speech codecs. The results indicated that processing the speech signal with the proposed irrelevancy removal technique before the encoding consistently improved the output quality. These promising results were attained without any optimisation of the original codecs. One of the most important future research objectives is the determination of the amount of the possible bit rate saving enabled when the proposed technique is used in the front end of an optimised speech codec.

TAMPEREEN TEKNILLINEN YLIOPISTO
Sähkötekniikan koulutusohjelma
Digitaali- ja tietokonetekniikan laitos
Lähdekorpi, Marja: Perceptual irrelevancy removal in narrowband speech coding
Diplomityö, 55 + 1 s.
Tarkastajat: Professori Jukka Saarinen (TTY)
DI Jani Nurminen (NRC)
Professori Ari Visa (TTY)
Rahoittajat: Tekes
Nokia Research Center (NRC)
Sähkötekniikan osasto
Huhtikuu 2003

Puhe on ihmisten välisessä viestinnässä hyvin oleellinen tiedonvälityskeino. Puheen avulla tapahtuva kommunikointi on tehokasta eikä se edellytä viestijöiden välistä näköyhteyttä. Puheen tärkeä asema viestinnässä heijastuu myös teknologian alalle, jossa kehitellään erilaisia puheenkäsittelyn menetelmiä mm. tallennus- ja tiedonsiirtosovelluksiin. Erityisesti langattoman viestinnän nopea yleistyminen on johtanut siihen, että yhtenä tärkeimmistä puheenkäsittelyn menetelmistä voidaan pitää puheenkoodausta. Sen avulla pienennetään digitaalisen puhesignaalin esittämiseen tarvittavaa bittinopeutta, jolloin myös puheen siirron vaatima kaistanleveys pienenee. Tallennussovelluksissa bittinopeuden alentaminen vähentää tallennuskapasiteetin tarvetta. Tehokkaille puheenkoodausmenetelmille on kysyntää monissa kaupallisissa sovelluksissa, joista tyypillisimmät liittyvät nykyajan vilkkaaseen matkapuhelinliikenteeseen.

Puheenkoodauksessa tavoitellaan yleensä alkuperäiseen esitykseen verrattuna suhteellisen alhaista bittinopeutta heikentämättä oleellisesti äänenlaatua. Ihmisen puheentuottomekanismin tunteminen tarjoaa tehokkaan apuvälineen tämän tavoitteen saavuttamiseksi, sillä äänihuulten synnyttämä heräte ja itse ääniväylä voidaan mallintaa erillisinä parametreinä. Mallinnuksen perustana on lineaarinen ennustus (linear prediction), jonka avulla voidaan karkeasti arvioida ääniväylän vaikutusta herätteeseen. Ääniväylää kuvaava informaatio koodataan lineaarisen ennustuksen kerrointen muodossa ja heräte mallintuu signaalina, jota kutsutaan residuaaliksi. Lineaarista ennustusta käytetään puhesignaalissa olevan redundanssin vähentämiseen lähes kaikissa olemassa olevissa puhekoodekeissa.

Tehokkaaseen puheenkoodaukseen pyrittäessä on tärkeää perehtyä paitsi puheen tuottamisen periaatteisiin, myös ihmisen kuuloaistin ominaisuuksiin. Psykoakustiikka eli kuulon psykofysiologia on mielekäs lähestymistapa kuuloaistin toiminnan tutkimiseen, sillä sitä soveltamalla vältetään hankalat ja kuuloelinten herkälle rakenteelle vaarallisetkin fysiologiset kokeet. Psykoakustiikan perusajatuksena on nimittäin tutkia korvaan saapuvan ääniärsykkeen ja sen aiheuttaman psyykkisen vasteen välistä yhteyttä. Psykoakustiikan tutkimustulosten perusteella kehitellään matemaattisia malleja kuuloaistimusten muodostumiselle ja näin kuulon toiminnan periaatteita päästään hyödyntämään signaalinkäsittelyn sovelluksissa. Alan tutkimusta on tehty jo vuosikymmenien ajan, mutta vasta suhteellisen uusissa tutkimusprojekteissa kiinnostus laajamittaiseen ihmiskuulon ominaisuuksien hyödyntämiseen on voimakkaasti kasvanut. Erityisesti audiosignaalien

koodaukseen on kehitetty erilaisia algoritmeja, joissa kuuloaistin toiminnan mallinnusta käytetään laajalti hyväksi. Puheenkäsittelyn puolella vastaavan idean soveltaminen on tähän asti ollut hieman vähäisempää.

Psykoakustiikan lainalaisuuksien hyödyntäminen puheenkoodauksessa voi ihannetapauksessa tarjota mahdollisuuden merkittävään bittinopeuden alentamiseen ilman havaittavaa äänenlaadun huonontumista. Vaihtoehtoisesti äänenlaatua voidaan parantaa bittinopeutta kasvattamatta. Koodaustehokkuuden paraneminen viittaa usein juuri tällaisiin saavutuksiin, vaikka kokonaistehokkuuteen vaikuttavat monet muutkin seikat, esimerkiksi laitteistolta vaadittava laskentakapasiteetti. Tässä työssä tarkastellaan mahdollisuuksia kasvattaa puheen koodaustehokkuutta puhesignaalia muokkaamalla. Tarkoituksena on vähentää signaalista kuuloaistimuksen kannalta epäoleellista informaatiota ennen varsinaista koodausprosessia ilman, että havaittu äänenlaatu kuitenkaan muuttuu. Menetelmän toimivuus perustuu maskaus- eli peittoilmiöön, jossa ääni voi peittää toisen, hiljaisemman äänen kuulumattomiin tai ainakin heikentää sen aistimista. Peittoilmiö on korvan rakenteesta ja toiminnasta johtuen kuuloaistin luontainen ominaisuus ja se vaikuttaa kuulohavaintoihimme kaikissa arkipäiväisissä tilanteissa. Äänisignaalin aiheuttaman peittoilmiön voimakkuutta eri taajuuksilla voidaan parhaiten arvioida peittokuulokynnyskäyrällä, jonka määrittämiseen käytetään psykoakustisia, ihmisen kuuloaistin kuvaamiseen kehitettyjä laskennallisia malleja.

Tässä työssä tutustuttiin muutamiin psykoakustisiin malleihin, jotka ovat puhe- ja audiosignaalien käsittelyssä merkittäviä ja yleisesti tunnettuja. Kaksi alunperin audiosignaaleille suunniteltua mallia sovitettiin kapeakaistaiselle puheelle, jotta niitä voitiin käyttää tutkittavan muokkausmenetelmän perustana. Kyseisessä menetelmässä puhetta käydään läpi lyhyinä, ajallisesti osittain päällekkäisinä kehyksinä, joista kullekin vuorollaan määritetään peittokuulokynnyskäyrä maskausmallin avulla. Kun näin saatua käyrää ja kehyksen spektriä verrataan toisiinsa, saadaan tunnistettua kehyksestä maskautuneet eli kuulon kannalta epäoleellisiksi tulkittavat taajuuskomponentit. Nämä taajuudet poistetaan käyttäen adaptiivista suodatusmenetelmää, jossa suotimen vaste muuttuu kehyksestä toiseen siirryttäessä puhesegmentin spektrin ja sitä vastaavan peittokuulokynnyskäyrän määräämällä tavalla.

Alustavan kuuntelun perusteella havaittiin, että toinen toteutetuista psykoakustisista malleista toimi työssä esitetyssä sovelluksessa huonosti. Kyseiseen malliin nojautuvaa muokkausmenetelmää ei testattu sen pidemmälle. Varsinaiset kokeet aloitettiin epävirallisella kuuntelutestillä, jossa kuuntelijoita pyydettiin vertaamaan alkuperäisiä ja työssä esitellyllä menetelmällä muokattuja suomenkielisiä testinäytteitä keskenään. Puheenmuokkausmenetelmän ei tulosten perusteella havaittu merkittävästi heikentävän äänenlaatua, mutta samalla objektiiviset mittaukset antoivat viitteitä siitä, että käsitelty puhe olisi mahdollista koodata hieman pienemmällä bittinopeudella kuin alkuperäinen. Koska menetelmä oli tarkoitettu puhekooderia edeltäväksi yleiskäyttöiseksi signaalinmuokkauslohkoksi, sitä testattiin myös kahden standardoidun, toimintaperiaatteeltaan erilaisen puhekoodekin yhteydessä. Toinen näistä oli puheen aaltomuodon approksimointiin perustuva CELP-koodekki (code-excited linear prediction) ja toinen oli tehokasta puheen parametrista mallinnusta hyödyntävä MELP (mixed excitation linear prediction) -koodausalgoritmi. Muokkauslohkon ja koodekin yhdistelmän tuottamaa subjektiivista äänenlaatua tutkittiin uuden kuuntelutestin avulla. Saatujen

tulosten mukaan molempien koodekkien ulostulon äänenlaatu parani hieman, kun sisäänmenosignaalina käytettiin alkuperäisen sijasta muokattua puhetta. Myös signaali-kohinasuhteeseen perustuvat mittaukset CELP-koodekin kanssa antoivat positiivisia tuloksia.

Näiden lupaavien tulosten perusteella voidaan alustavasti päätellä, että työssä kehitetyn kuuloaistin ominaisuuksia hyödyntävän puheenmuokkausmenetelmän käyttäminen kapeakaistaisessa puheenkoodauksessa voi parantaa koodaustehokkuutta. Koodekkien yhteydessä saavutettujen tulosten merkittävyyttä lisää se, että kumpaakaan koodekkia ei ollut millään tavalla optimoitu muokatun puhesignaalin käsittelemistä varten. Vielä positiivisempia tuloksia todennäköisesti saavutettaisiin koodekkien asianmukaisen optimoinnin jälkeen. Jo saavutetut, suuntaa antavat tulokset kuitenkin antavat aiheen uskoa, että esitettyä menetelmää kannattaa kehittää tulevaisuudessa.

Tämä diplomityö tehtiin Tampereen teknillisen yliopiston digitaali- ja tietokonetekniikan laitoksella osana laajempaa matalan bittinopeuden puheenkoodaukseen keskittyvää tutkimusprojektia. Kyseinen projekti kuului Teknologian kehittämiskeskuksen (Tekes) USIX – Uusi käyttäjäkeskeinen tietotekniikka -teknologiaohjelmaan. Yhteistyö-kumppanina toimi Nokia Research Center (NRC). Kirjoittajan oma osuus tästä työstä käsittää kaikki esitetyt teoriaan liittyvät taustaselvitykset, toteutukset Matlab-ohjelmistolla, menetelmän kehittelyn ja suurimman osan sen testauksesta sekä dokumentoinnin. Kuuntelutestien järjestämiseen saatiin teknistä apua muilta projektiin osallistuneilta henkilöiltä. Tärkeimpänä oman toteutuksen ulkopuolisena kokonaisuutena mainittakoon puhekoodekkien tuottamaa äänenlaatua selvittävä kuuntelutesti, joka suoritettiin NRC:llä.

List of symbols

General symbols

F_S	sampling frequency
f	frequency in hertz
i	critical band index
k	FFT index
L_S	sound pressure level
N_F	frame length in samples
n	sample index
p	sound pressure
p_0	reference sound pressure, 20 µPa
$q(f)$	absolute threshold of hearing
$X(k)$	complex spectrum of windowed speech segment
$z(f)$	critical band number in Bark

Masking model of Johnston

$\mathbf{b}$	critical band spectrum
$b(i)$	energy of one critical band
β	coefficient of tonality
$\mathbf{c}$	spread critical band spectrum
$c(i)$	spread energy of one critical band
Im	imaginary part
i_{max}	index of the highest critical band
j	Bark index of masker signal
λ_{dB}	spectral flatness measure in decibels
λ_{dBmin}	minimum value of spectral flatness measure in decibels
$o(i)$	offset in decibels for the ith critical band
$P(k)$	power spectrum
Re	real part
$\mathbf{S}$	spreading function matrix
$\sigma_l(i)$	lower boundary FFT index of the ith critical band
$\sigma_u(i)$	upper boundary FFT index of the ith critical band
$s_{dB}(x)$	spreading function in decibels
$s(x)$	spreading function on linear scale
$\mathbf{t'}$	renormalised masking threshold
$t(i)$	spread masking threshold
$t'(i)$	renormalised masking threshold of one critical band
x	distance in Bark between masker and maskee, $\lvert j - i \rvert$

Masking model of PEAQ

A_{max}	maximum amplitude of test sine
$A(j,L)$	gain normalisation function for spread of energy
$\alpha(i)$	control parameter for time domain smearing
$B(f)$	hertz to Bark conversion function
$B^{-1}(z)$	Bark to hertz conversion function
$B_S(i)$	level normalisation factor for spread of energy
$\gamma(f_x)$	normalisation factor related to test sine
Δz	critical band resolution
$E(i,n_f)$	excitation pattern
$E_f(i,n_f)$	filtered (smeared) excitation pattern
$E_S(i)$	unsmeared excitation pattern
$E_S(i,n_f)$	unsmeared excitation pattern with frame index
F_{ss}	frame rate
$f_c(i)$	centre frequency (in hertz) of the ith frequency band
$f_l(i)$	lower edge (in hertz) of the ith frequency band
$f_u(i)$	upper edge (in hertz) of the ith frequency band
f_x	frequency of test sine, 1019.5 Hz
G	combined scaling factor for input signal
g	input level scaling factor
j	critical band index
$k_l(i)$	lowest FFT index belonging to the ith frequency band
$k_u(i)$	highest FFT index belonging to the ith frequency band
κ	normalisation factor in the input level adjustment
$L(i)$	soud pressure level of pitch pattern in the ith band
L_p	assumed sound pressure level of a full scale test sine
$M(i)$	masking threshold
$m(i)$	weighting factor for excitation patterns
N_C	number of critical bands
n_f	frame index
$P_e(i)$	energy in the ith band
$S(i,j,L)$	spreading function
$S_{dBl}(i,j)$	lower slope of spreading function in decibels
$S_{dBu}(i,j,L)$	upper slope of spreading function in decibels
$\tau(i)$	time constant of time domain smearing filter
τ_{min}	minimum time constant of time domain smearing filter
τ_{100}	time constant of time domain smearing filter at 100 Hz
$U(i,k)$	contribution from the kth frequency component to the ith band
$W(k)$	outer and middle ear frequency response at discrete frequencies
$W_{dB}(f)$	outer and middle ear frequency response in decibels
$W_f(f)$	outer and middle ear frequency response on linear scale
$X_W(k)$	weighted and scaled power spectrum of input speech segment

List of abbreviations

AAC	Advanced Audio Coding
ACR	absolute category rating
AMR	adaptive multi-rate
BM	basilar membrane
CCR	comparison category rating
CELP	code-excited linear prediction
CF	characteristic frequency
DFT	discrete Fourier transform
DPSS	discrete prolate spheroidal sequence
ERB	equivalent rectangular bandwidth
FFT	fast Fourier transform
IEC	International Electrotechnical Commission
IFFT	inverse fast Fourier transform
IHC	inner hair cells
IRS	intermediate reference system
ISO	International Organization for Standardization
ITU	International Telecommunications Union
LP	linear prediction
LPC	linear predictive coding
LSF	line spectral frequency
MELP	mixed excitation linear prediction
MNRU	modulated noise reference unit
MOS	mean opinion score
MOV	model output variable
MPEG	Moving Picture Experts Group
OHC	outer hair cells
PEAQ	perceptual evaluation of audio quality
PXFM	perceptual transform coder
SMR	signal-to-mask ratio
SNR	signal-to-noise ratio
SPL	sound pressure level
WI	waveform interpolation

1 Introduction

Speech compression techniques have an important status in the modern communication systems. In the transmission and storage applications where speech signals are usually in a digital form, the bit rate, i.e., the number of bits required to present one second of speech, is one of the most salient attributes of the representation. The bit rate controls the bandwidth needed for transmission and determines the memory requirements of the storage systems. Consequently, speech coding is needed as a means to obtain a more compact representation of the signal. Especially in wireless communication where the data transfer rate often forms a constraint, speech coding plays an important role. Several commercial applications of the compression techniques can be found, starting from cellular phones and ending up in the possibilities of multimedia presentations. Naturally, the technological advances also constantly stimulate the emergence of new applications.

Speech coding is comprised of encoding and decoding. The encoder produces a low-rate bit stream from the speech signal while the decoder reconstructs the signal with certain accuracy based on the contents of the received bit stream. The main objective in speech coding is a relatively low bit rate compared to the uncompressed representation of the signal, yet avoiding significant degradation of the speech quality. A case apart are the military applications where the good quality of the signal is more easily traded off against a very low bit rate. In contrast to audio signals, the coding of speech is considerably facilitated by efficient source models by which the vocal tract and its excitation signal can be described with separate parameters. An important factor increasing the efficiency in source modelling is the reduction of redundancy by the linear prediction (LP) technique that forms the basis of almost all modern speech codecs.

In the pursuit of the best possible speech quality within the available bit rate, it should be noted that the most important speech quality criterion is usually that judged by humans according to the perception of the sound. Consequently, it is very advantageous to develop speech coding methods with an eye to minimising the perceptually disturbing distortions in the coded speech (i.e., in the output of the decoder). This creates the need for the modelling of the human auditory system. Due to the highly sophisticated structure and elaborate operation of the auditory system, the research within this field is very challenging. Investigations into the anatomy and physiology of the auditory system have been carried out over several decades. Despite the remarkable progress that has been made

in understanding the complex mechanisms of hearing, all details about the human hearing—especially the top-level operations in the brain in connection with information processing—are still not completely understood. A very interesting and useful approach in the research is provided by psychoacoustics. It is a science that examines the relationship between sounds and their psychological effects. Connections are sought between the acoustic stimuli and the resulting sensations without the need for physiological experiments that are often laborious and also risky for the normal operation of the hearing. The results of the psychoacoustic research form the basis of the work presented in this thesis.

The most salient psychoacoustic phenomenon utilised in speech processing is masking. By the definition of the American Standards Association, rewritten in [25, p. 89], masking is "the process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound". Masking reflects the limited frequency selectivity of the auditory system; the strongest effects can be observed when a tone is masked by noise whose centre frequency is equal to the frequency of the tone. In addition to the relative frequencies and the tonelike or noiselike nature of the sounds, the relative levels and the temporal positions have an impact on the total masking effect. An approximate measure of the amount of masking can be obtained by evaluating a masking threshold. It indicates the sound pressure level at which a test sound is just audible in the presence of a masking sound, i.e., a masker [43]. For complex signals such as speech, the exact evaluation of the masking threshold is practically impossible and coarse simplifications must be made, but even the approximate threshold gives valuable information for many applications.

The utilisation of the psychoacoustic principles in signal compression has been under increasingly active research during the recent years. However, it is not by any means a novel idea. A simple structure exploiting the masking phenomena is the traditional noise shaping weighting filter in the coding of the LP residual that has been used for decades [1]. Moreover, in 1979, Schroeder *et al.* have proposed a technique for objectively measuring the quality of the coded speech, yet based on perceptually relevant criteria [33]. The work published by Schroeder has been cited in the description of another quality evaluation procedure, the ITU standard for perceptual evaluation of audio quality (PEAQ) [9]. Parts of Schroeder's work have also been utilised in the auditory masking model proposed by Johnston, designed to control the quantisation stage of an audio coder [15]. Another audio application has been reported in [29] where very promising results have been achieved by incorporating the auditory modelling into the sinusoidal analysis-synthesis of audio signals. In the field of speech signal processing, one of the most noticeable applications of masking models has been in the speech enhancement [37 – 40, 42]. Especially worth mentioning, however, is the work published in [19 – 22] since it has been motivating the work within this thesis. In that work, simultaneous masking has been incorporated in the linear prediction, yielding improvements in the perceptual quality of coded speech without increasing the bit rate. The basic idea in the method is to calculate the linear prediction coefficients using a speech signal from which the simultaneously masked frequency components have been removed. The remaining parts of the coding process use the original speech signal.

The purpose of this thesis is to examine how the masking phenomenon can be further exploited in speech coding. Unlike many of the previous applications that have

concentrated on suppressing the quantisation noise by the useful signal [15, 26] or tried to extract the perceptually relevant components to be coded [19 – 22, 29], this work aims at a generic method of reducing the perceptual irrelevancy of the speech signal before it is fed into the encoder. Those components of the speech signal that, due to the masking phenomenon, cannot be perceived by a human listener are detected using an auditory model and removed by adaptive filtering. The resulting signal is to be used in all phases of the encoding. Since the irrelevancy removal procedure is placed in front of a speech coder, it will be referred to as preprocessing.

The aim in the proposed preprocessing method is to make the speech signal applicable for more efficient coding without degrading its perceptual quality. The compressibility and quality of the preprocessed speech are first examined outside a codec and then in connection with two standardised speech codecs, an adaptive multi-rate (AMR) code-excited linear prediction (CELP) codec [5] and a 2.4 kb/s mixed excitation linear prediction (MELP) codec [34]. Preliminary results are obtained about the ability of the codecs to operate with preprocessed speech signals. Since optimising the codecs for the preprocessed speech fell outside the scope of this thesis, the emphasis is mainly on the perceptual quality of the speech rather than on the extent of the bit savings enabled by this technique.

The work within the perceptual preprocessor presented here is a part of a larger project directed at the research of low bit rate speech coding. However, the work documented in this thesis has mainly been done by the author. The most significant exception is the listening test incorporating the speech codecs; it has been organised by the Nokia Research Center using the preprocessor provided by the author. Other smaller parts involving technical assistance from outside will be indicated in the text. The implementation as well as most of the test procedures have been written using Matlab.

This thesis is organised as follows. The basic psychoacoustic theory behind this work is presented in Chapter 2, discussing both the structure and the operation of the human auditory system. Chapter 3 continues with the theory by giving detailed descriptions of the two most essential masking models in this research work and by also providing a view on some other existing models. In Chapter 4, one of the auditory models is chosen as the basis of the preprocessor and necessary modifications to the model are proposed to make it work sensibly in the application where a telephone bandwidth (0.2–3.4 kHz) is used. In addition to describing the implementation of the perceptual irrelevancy removal technique, Chapter 4 considers some other possible applications that could utilise the masking threshold calculated within the preprocessor. The evaluation of the proposed method starts in Chapter 5, where the detailed descriptions of the test procedures together with the results are presented. Chapter 6 discusses the results and considers some development ideas related to the method. Finally, conclusions and future research possibilities are outlined in Chapter 7.

2 Human hearing

The human auditory system is a highly complicated mechanism. It is able to detect extremely small variations in the sound pressure and to convert them into meaningful auditory sensations. Even a sound pressure fluctuation that has a magnitude of only a tenth millionth part of the atmospheric pressure is sufficient to be perceived by the human ear [43]. The energy of the sound waves captured by the ear lobe travels to the inner ear and goes through frequency analysis which is generally assumed to be affected by very sophisticated active and nonlinear processes. In the final conception of the sound, cognitive effects also play a role. During the last three decades, remarkable progress has been made within the research of the human hearing but many details still need to be explained. The field of psychoacoustics has an interesting standpoint in the research since it examines directly the relationships between acoustic stimuli and the associated sensations.

In the study of the human auditory system, it is a normal practice to make a distinction between the peripheral part, and the part that contains the nervous system and leads to the final auditory sensation. The peripheral part of the human auditory system refers to the elements in which the oscillations due to a sound stimulus retain their original character. Zwicker designates the function of the peripheral system as preprocessing of sound [43]. In contrast, the neural processing takes place in the second region of the hearing system that consists of the auditory sensation area of the brain together with the nerve fibres [43]. However, the nervous system and its functions are in excess of the scope of this thesis and they are not covered here. Rather, emphasis is placed on the structure and operation of the peripheral part, as the aim is to introduce the properties influencing the perception of sounds. This will serve as a basis for the auditory models, some of which will be presented in Chapter 3.

The frequency analysis performed by the ear has a certain finite resolution, leading to the concept of masking. It is a phenomenon in which one sound can drown out another sound either partially or totally. The relative levels and frequencies of these sounds determine for the most part the degree of masking, but temporal factors also have some influence. The quantitative effect of masking can be depicted by means of a masking threshold. It shows

the sound pressure level that a test tone must have in order to be just audible in the presence of a masker [43]. The masking threshold is one of the main concepts in this thesis and it is viewed more precisely and mathematically in Chapter 3.

This chapter deals with the main points of the human auditory system. Section 2.1 contains an overview of the peripheral part of the auditory system, considering both physical structure and operation. Section 2.2 concentrates on the capabilities and constraints of the human hearing, starting with the limits of the hearing area. After that, the limited frequency resolving capabilities of the auditory system and the associated masking phenomena are treated. While even specialists do not yet understand all the details about the masking process, the basic concepts of the auditory filter and critical band are presented. Finally, the masking effects in both frequency and time domain are discussed. Section 2.3 contains a brief summary of the chapter.

2.1 Structure of human ear

The purpose of the ear is to capture sound waves and to convert the acoustic energy of these small pressure fluctuations into electrical nerve impulses. The nerve fibres convey the information to the brain in which it is perceived as sounds. Reciprocally, the brain sends information to the ear, thus actively controlling some of the functions of the so-called sound preprocessing [25]. The ear also contains the vestibular organ that contributes to balancing the body, but it is not studied here because it has no effect in the perception of sounds.

A simplified representation of the structure of the human ear is shown in Figure 2.1. It can be examined in three parts: the outer, middle and inner ear. In the following three subsections, each of these parts will be considered separately.

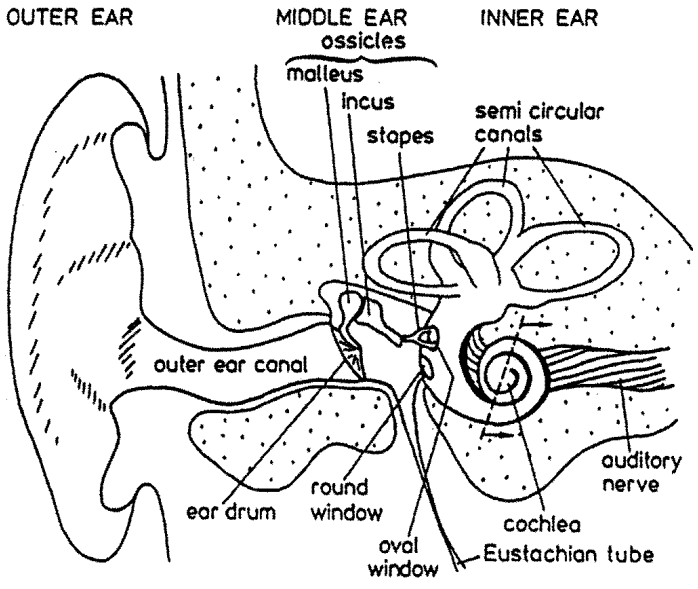

Figure 2.1. Structure of the human ear. The figure has been adopted from [43].

2.1.1 Outer ear

The outer ear is composed of the *pinna* (the visible part, ear lobe), the *meatus* (auditory canal) and the *tympanic membrane* (eardrum). The pinna collects the sound energy which then travels down the meatus and makes the eardrum vibrate. The eardrum is a hermetic membrane whose function is to convey the acoustical energy to the middle ear.

The pinna and the outer ear canal have a strong influence on the incoming sound. The pinna filters the sound in a way that depends on its inlet angle, thus providing cues to the localisation of sound. This works especially at high frequencies where the shadowing effect of the pinna attenuates the sounds that come from behind the listener. At low frequencies, this does not take effect because the wavelengths are too large compared to the dimensions of the pinna [24]. The meatus, acting like an open pipe with a length of approximately 2 cm, has a resonant frequency at about 4 kHz. Consequently, the meatus is responsible for the high sensitivity of hearing around this frequency. In addition to the sound wave modifications performed by the pinna and the meatus, the head and shoulders of the subject have the effect of shadowing and reflection [43].

2.1.2 Middle ear

The middle ear is a chamber that contains the auditory ossicles: *malleus* (hammer), *incus* (anvil) and *stapes* (stirrup), the smallest bones in the human body. The hammer is firmly attached to the eardrum and the stirrup touches the oval window which serves as the interface to the inner ear. Figure 2.1 also shows the *Eustachian tube* that forms a connection from the middle ear to the pharynx. It can be opened briefly by swallowing and it is used for equalising the pressure on the different sides of the eardrum so that the ossicles are able to serve the energy transfer with maximum efficiency [43]. The middle ear provides the two important functions of impedance transformation and amplitude limiting, both of which will be briefly described next [31].

The oscillations of air particles in the outer ear need to be converted into motions of the water-like fluid in the inner ear in order to excite the auditory sensory cells. The middle ear performs impedance transformation to ensure the efficient transfer of the acoustical energy, avoiding large reflections. The impedance transformation is based on both the lever system constructed from the ossicles and the ratio of the area of the eardrum to that of the small oval window. [31, 43]

The amplitude limiting is made possible by the tiny inner ear muscles that are attached to the auditory ossicles. When the subject is exposed to very intense sounds, these muscles contract and thereby limit the transmission of sound through the ossicles. This operation, called the middle ear reflex, may help to protect the vulnerable structure of the inner ear, though it only works at low frequencies. The reflex also decreases the audibility of self-generated sounds by getting activated at the starting time of vocalisation. [25]

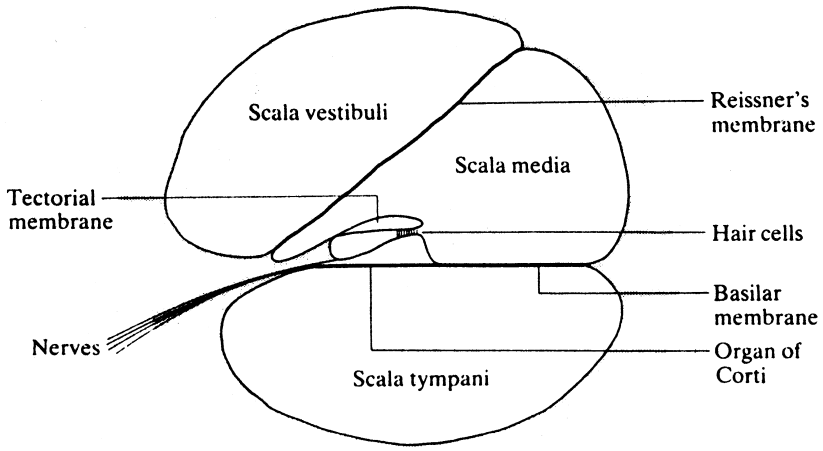

Figure 2.2. Cross-section of the cochlea [31].

2.1.3 Inner ear

The inner ear, comprising of the *cochlea* and the semi-circular canals of the *vestibular organ*, is the most complicated part of the ear. The cochlea transforms the mechanical oscillations at the oval window into electrical impulses to be detected by the auditory nerve. The structure of the cochlea is shown essentially in Figure 2.2.

The shape of the cochlea resembles a snail and it is filled with nearly incompressible fluids and surrounded by very hard bone. The cochlea is divided along its length into three channels by the *basilar membrane* (BM) and the *Reissner's membrane*, as seen in Figure 2.2. A small opening, called the *helicotrema*, connects the *scala vestibuli* and *scala tympani* at the far end of the cochlea, the *apex*. A surface wave disturbance, caused to the fluid by the inward movement of the oval window, travels through the scala vestibuli and the helicotrema and then back along the scala tympani. At the base of the cochlea, the scala tympani ends at another opening, the round window, which enables mechanical pressure release for the cochlea. [24]

The incoming sound, moving the oval window, results in longitudinal displacement of the fluids inside the cochlea and also in vertical displacement of BM. The position of the maximum displacement of BM depends on the frequency of the stimulus; low frequencies produce strongest oscillations near the apex and high frequencies near the oval window. The cochlea thus performs the very important function of analysing the frequency content of the incoming sound. The frequency of the stimulus that causes maximum response at a given point on BM is called the characteristic frequency (CF) for that point. However, the situation gets somewhat complicated with other than pure sinusoidal signals. If two frequency components of the stimulus are sufficiently close to each other, BM fails in the exact frequency-to-place conversion and only a single, broader maximum can be observed in the BM response. This phenomenon is almost certainly involved with masking which is discussed in Section 2.2. [25]

The vibrations of BM are detected by the hair cells of the *organ of Corti* which lies on BM. It contains two different kinds of hair cells, the inner hair cells (IHC) and outer hair cells (OHC), both having their own special functions. It seems that IHC convey most, or

even all, information about the sounds to the brain. OHC receive messages from the brain through several *efferent* (descending) nerve fibres. These messages are most likely used for active processes affecting the high sensitivity (discussed in Subsection 2.2.1) and sharp tuning of BM. [25]

2.2 Properties of hearing

The concept of hearing area refers to the ranges of frequency and sound pressure values within which the human ear generally perceives sound. Reviewing the limits of the hearing area is the first step in studying the properties of the human hearing. Another very commonly discussed property of the auditory system is the masking phenomenon. Masking is a process in which the threshold of audibility of a sound is raised due to the presence of another sound. These two sounds are referred to as the maskee and the masker, the latter representing the one that causes the shift in the threshold. Masking as a whole is a very complicated phenomenon, containing some details that are still not completely understood. However, masking effects can be experienced in everyday life. For example, it is more difficult to hear what the person in the next room is saying when the television is blaring out, compared to the situation in which the interfering sounds from the television are muted. The reason for not hearing what the person tries to tell is most often based on the inevitable masking property of the hearing system and not the lack of interest or concentration. In other words, one cannot hear the talking even by trying harder unless the television is turned down.

In order to be able to evaluate which parts of the incoming sound are masked, the main points of the operation of the human hearing are presented in this section. The following three subsections introduce the limits of the hearing area and the important concepts of auditory filter and critical band. After that, the simultaneous and temporal masking effects are discussed. A more detailed description of these subject matters can be found in [24] and [25], on which the theory presented in this section is mainly based.

2.2.1 Hearing area

As discussed earlier, the human ear is very sensitive to sound stimuli. While the atmospheric pressure at the sea level is about 10^5 pascals (Pa), the relevant sound pressure fluctuations from the standpoint of hearing lie between 10^{-5} Pa and 10^2 Pa [43]. The upper limit denotes the threshold of pain, but it must be noted that the limit of damage risk lies considerably lower. Usually, this wide range of pressures, p, is controlled by the term sound pressure level (SPL), L_S, which is defined by the equation

$$L_S = 20 \log\left(\frac{p}{p_0}\right) \text{ dB}, \tag{2.1}$$

where the reference value p_0 is standardised to 20 μPa [43]. The perception stimulated by a certain sound pressure level depends strongly on frequency. The audible frequency range

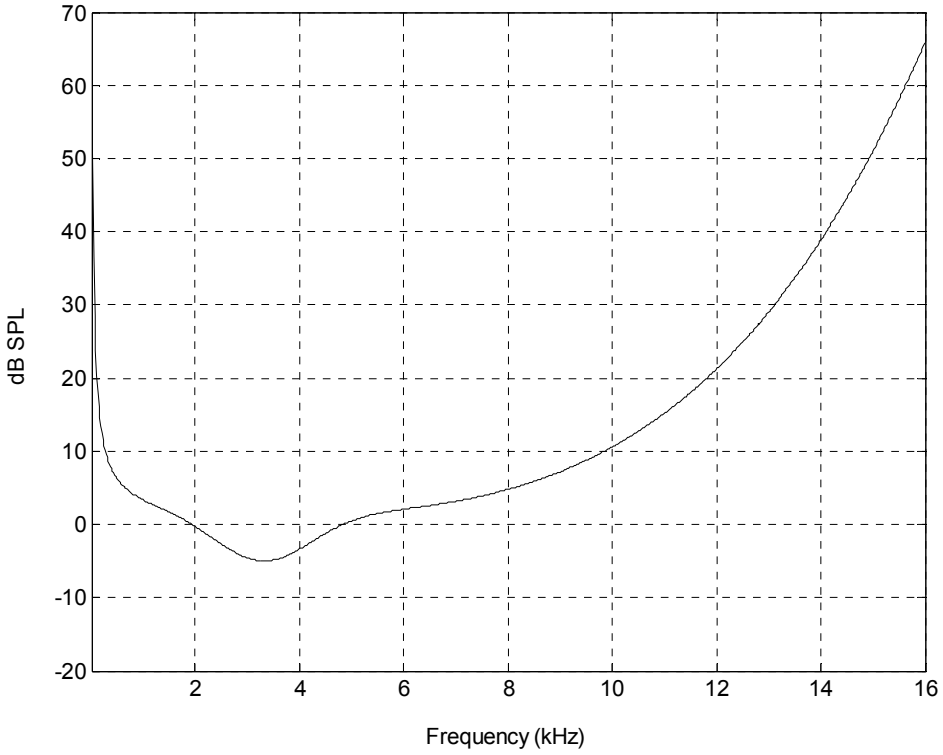

Figure 2.3. Typical absolute threshold of hearing as a function of frequency.

is from about 20 Hz to 16 kHz but the limits can be even 16 Hz and 20 kHz for a very keen ear. For the elderly, the upper frequency limit can drop to as low as 10 kHz [31].

The absolute threshold of hearing, also known as the threshold in quiet, signifies the minimum sound pressure level of a pure tone that is enough for the tone to be audible in the absence of any interfering voices, i.e., in quiet. A widely known approximation for the absolute threshold of hearing is that given by Terhardt [35],

$$q(f) = 3.64 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 10^{-3} f^4 \text{ dB SPL}, \tag{2.2}$$

where the frequency, f, is given in kHz. The shape and vertical position of the threshold curve vary somewhat between subjects but, for normal hearing, Equation (2.2) represents a typical curve. It is shown as a function of frequency in Figure 2.3. A curve like this can be obtained by averaging many individual absolute thresholds measured in many subjects. The dip in the absolute threshold curve in the neighbourhood of 4 kHz indicates the high sensitivity of hearing and also the high susceptibility to hearing impairment in this region [43]. It results from the meatus resonance mentioned in Section 2.1.1.

2.2.2 Auditory filters

A prominent contributor to the idea of auditory filters was Fletcher who, back in 1940, measured the detection threshold of a sinusoidal signal in the presence of a bandpass noise masker. In his experiment, the noise power density was held constant and its centre frequency was always the same as the signal frequency. As the noise bandwidth was increased, the threshold of the signal also increased at first, but after a certain noise bandwidth had been achieved, the signal threshold levelled off. The results of Fletcher's experiments led to a set of assumptions known as the power spectrum model. Although it involves slight inaccuracies, it gives a usable aspect to the masking process. [25]

Basically, the power spectrum model suggests that the peripheral auditory system contains a bank of linear overlapping bandpass filters called auditory filters. It is assumed that when trying to detect a signal in a noisy environment, only one filter whose centre frequency is close to the frequency of the signal is being used. According to the model, this auditory filter blocks out most of the noise and only the part passing through the filter affects the masking of the signal. In reality, the perception of complex signals, e.g. speech, depends on the outputs of several auditory filters and not just one. The assumption of linear auditory filters is also incorrect since, strictly speaking, the shape of the filter changes slightly with the input level. As the level of the stimulus is increased, the slope on the low-frequency side of the auditory filter becomes less sharp while the high-frequency skirt becomes steeper. This is illustrated in Figure 2.4. The bandwidth of auditory filters is treated in the next subsection. [25]

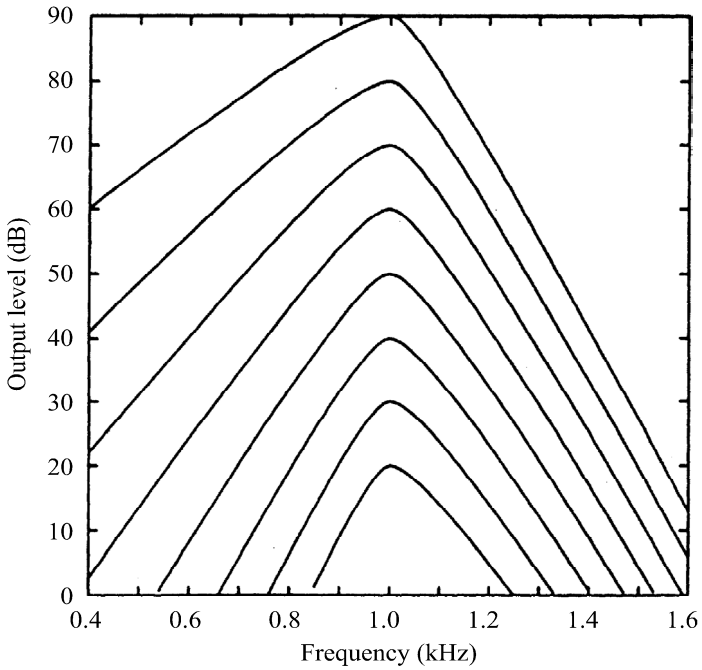

Figure 2.4. Shape of the auditory filter, adopted from [24].

2.2.3 Critical bands and Bark scale

In his band-widening experiment, Fletcher introduced the concept of critical bandwidth (CB), denoting the noise bandwidth limit at which the detection threshold of the signal (tone) ceased to increase. For simplicity, he thought that the auditory filter could be approximated as having a rectangular shape and a passband width equal to CB. The shape of the auditory filter is not really rectangular, as Fletcher also knew, but this kind of a model can be useful in evaluating the masking effects in many applications such as those presented in this thesis. Fletcher suggested that, with this rectangular model, CB could be evaluated by measuring the threshold of a sinusoidal signal in broadband white noise [25]. In this method, the power of the tone and the power spectral density of the noise masker are first measured. The noise power within the same critical band with the signal is then equal to the product of the measured power spectral density and the CB of the band in question. Fletcher's suggestion was that CB would simply be such that the ratio of the signal power to the noise power inside the critical band is equal to unity. In the described conditions, the tone would be just masked by the noise. Zwicker and Fastl have later presented several methods for determining the CB values [43] and concluded that the threshold is reached when the ratio of the signal power to the power of the noise lies between 0.25 and 0.5.

The critical bandwidth can also be explained based on the physical structure of the inner ear. Each point on the basilar membrane (BM) responds only to a certain range of frequencies, which leads to the idea that these different points correspond to auditory filters with different centre frequencies [25]. When the bandwidth of the auditory filter is expressed by means of the equivalent rectangular bandwidth (ERB), the relation to BM is very simple: “each ERB corresponds to a distance of about 0.89 mm on the basilar membrane [24, p. 175]”. ERB is defined so that the power of a signal inside the rectangular band equals the power of the same signal in the passband of the real filter. The bandwidth of the auditory filter, and hence also the ERB value, increases with increasing centre frequency.

A commonly used scale for signifying the critical bands is the Bark scale that divides the audible frequency range of 16 kHz into 24 abutting bands [43]. Figure 2.5 illustrates the relationship between the frequency in hertz and the critical-band rate in Bark, both in proportion to the length of the unwound cochlea. The characteristic frequencies related to the points of BM can also be roughly estimated from the figure. An approximate analytical expression to describe the conversion from linear frequency, f, into the critical band number z (in Bark) is

$$z(f) = 13\arctan(0.76f) + 3.5\arctan\left[\left(\frac{f}{7.5}\right)^2\right] \tag{2.3}$$

and the critical bandwidth (in Hz) for a given centre frequency can be evaluated by

$$BW(f) = 25 + 75\left(1 + 1.4f^2\right)^{0.69}. \tag{2.4}$$

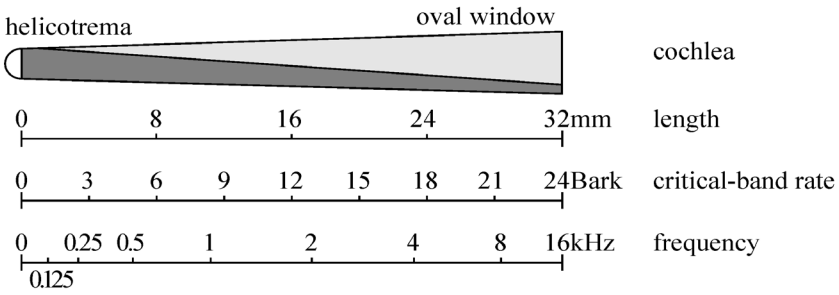

Figure 2.5. Frequency in hertz and the critical-band rate scale relative to the length of the unwound cochlea (after [43]).

In the presented equations, adopted from [43], f is given in kHz. By the definition of the Bark scale, each critical band has a width of one Bark. This property conveniently simplifies, among other things, the calculation of the spread of masking, i.e., the effect of adjacent critical bands on the amount of masking in a particular band. A list of the critical bands, as presented in [43], is given in Appendix A. The numerical values shown in the appendix do not exactly match with those given by Equations (2.3) and (2.4) since the equations can only present approximations to the unit conversion and the bandwidth evaluation. It should also be noted that the presentation of the critical bands as discrete abutting bands is just a coarse simplification. In reality, the number of overlapping critical bands is indefinite and they form a continuous series [25].

2.2.4 Simultaneous masking

In simultaneous masking, the masker and the maskee are present at the same time. Masking occurs when the basilar membrane fails to separate these two sounds and thus, the masking phenomenon can be considered to reflect the limits of the resolution of the frequency analysis performed by the inner ear. The amount of masking depends on the levels, frequencies and durations of the masker and the maskee, and essentially on the nature of the masker. The strongest effects can be observed when a tone is masked by noise whose centre frequency is equal to the frequency of the tone. Figure 2.6 represents the masking threshold of a test tone in a situation just described. It shows the dependence of the masking threshold on the level of a noise with a critical bandwidth, i.e., a bandwidth of about 160 Hz as the centre frequency is 1 kHz. The maximum of each masking threshold curve is only 3 dB below the level of the masker.

Compared to noise, tone-like sounds are considerably less efficient in masking and their behaviour as maskers is far more difficult to measure. The listeners, especially inexperienced ones, often hear beats which result from the frequency difference between the test tone and the masker tone. Partly because of this, the listeners report a perception of a sound in addition to the masker even though the test tone itself cannot be heard [43]. In order to render a test tone (and the difference tones) or noise inaudible, the level difference of 3 dB mentioned above for noise maskers is not adequate in the case where the masker has a tonal nature.

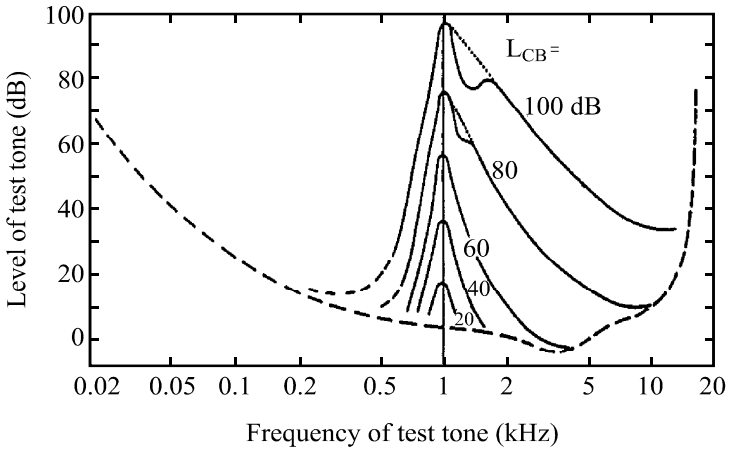

Figure 2.6. Masked threshold for a 1 kHz tone masked by a critical-band wide noise [43].

The dips in the threshold curves of Figure 2.6 for the two highest masker levels are explained in [43] to be a consequence of nonlinear effects in the auditory system. Another essential feature in Figure 2.6 is the nonlinear increase in the upper slope of the masking threshold with increasing masker level. In comparison, the auditory filter shows this kind of behaviour on its low-frequency side (see Figure 2.4). The upward spread of masking can be viewed through the concept of excitation pattern, the shape of which the masking threshold curve roughly indicates. The excitation pattern can be interpreted as an internal representation of the spectrum of the sound, i.e., a representation of the amount of activity evoked by a sound as a function of the CF of the excited neurone [25]. Since the upper slope of the excitation pattern, and hence also that of the masking pattern, is determined by the lower part of the auditory filter and vice versa, the spread of masking towards upper frequencies occurs. Details about the derivation of the shape of the excitation pattern from the auditory filters can be found in [25]. Methods for calculating the spread of masking in the Bark domain for a practical application will be presented in Chapter 3.

2.2.5 Non-simultaneous masking

Masking can also occur when the masker and the maskee are presented consecutively in time, without any overlapping time section. This phenomenon, called non-simultaneous masking, is even more poorly understood than simultaneous masking and it is often considered to be of secondary importance when the masking effects are estimated on a coarse level. Non-simultaneous masking, also known as temporal masking, is typically divided into two different cases: backward and forward masking. In the former, the maskee appears before the masker and in the latter, the temporal positions are reversed. The terms prestimulatory and poststimulatory masking are also sometimes used [24]. Figure 2.7 shows the relevant time scale for masking effects with a rather long masker duration of 200 ms. The amount of forward masking depends strongly on the duration of the masker, shorter duration resulting in the masking threshold to drop faster [43]. The ordinate in the figure, labelled as the sensation level, represents the level above the absolute threshold of the masked sound and not its sound pressure level. Thus, when the masking threshold curve falls to the sensation level of 0 dB, the effect of temporal masking ceases.

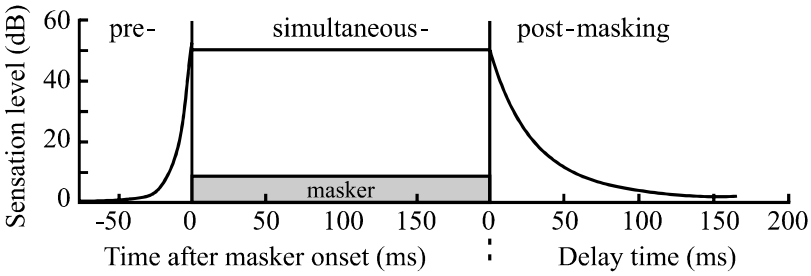

Figure 2.7. Premasking, simultaneous masking and postmasking relative to the masker (after [43]).

Trained listeners often show considerably less backward masking than those who have got little or no practice [25]. In any event, the effect of backward masking is minor and for that reason it will not be discussed further. A typical example of forward masking in speech is the situation in which a plosive follows a loud vowel and gets masked. This is a very customary situation to which the human communication has been adapted, and therefore, it does not usually hinder intelligibility. However, this kind of a forward masking phenomenon may be advantageous from the viewpoint of speech coding. Apart from the masker duration, the level and bandwidth of the masker, as well as the frequencies of the maskee and the masker also influence the forward masking. Moore has gathered the most important properties of forward masking into a list [25, p. 129] but it is not examined here. For the scope of this thesis, it is sufficient to be aware of the temporal spreading of the masking pattern.

2.3 Summary

In this chapter, the human auditory system was examined from both structural and operational point of view. The emphasis was on the somewhat simpler peripheral part, in contrast to the complicated mechanisms of the nervous system. The three parts of the peripheral auditory system were described separately and their capability to process sounds was presented as the limits of the hearing area. Next, the auditory filters and their bandwidths were presented as a way to model the frequency analysis that takes place in the inner ear. The Bark scale was introduced to allow simple division of the audible frequency range into critical bands. Finally, both simultaneous and non-simultaneous masking phenomena were treated briefly. Although the facts introduced in this chapter only merely scratched the surface of the research into human hearing, they will be useful for the following chapter in which some mathematical auditory models will be presented.

3 Psychoacoustic models

The need for masking models arises from the objective of developing speech and audio codecs that preserve a good perceptual quality of the output signal despite significant reduction of bit rate. Masking models are utilised, for example, for shaping the noise introduced in the coding process such that it is masked as effectively as possible by the signal of interest, e.g. speech. This way the perceptual quality can be remained high even though objective measures, such as the mean squared difference between the input and output of the codec, may imply rather remarkable changes in the signal.

Utilising the human auditory properties and the derived masking models is not a new idea. Schroeder reported already in 1979 his method of exploiting the auditory masking effects in speech coders [33]. His approach contained an auditory model that was used to evaluate the loudness of the quantisation noise and that of the signal, providing an objective measure of speech signal degradation caused by the coder. Two more recent masking models, both utilising parts of the work done by Schroeder, are presented in this chapter. The first of them has been announced by J. D. Johnston [15] and the other is a part of the standard ITU-R BS.1387, also known as the perceptual evaluation of audio quality (PEAQ) [9]. Many other models have also been published, for example by the Moving Picture Experts Group (MPEG), but they are examined only coarsely here because they are not so salient for this work. Rather, the emphasis is on the two masking models that have also been implemented in Matlab during this project. Both of the presented models have originally been designed for audio signals and provide a way to calculate the masking threshold that was introduced in the beginning of Chapter 2. However, some properties are very different from one model to another. Johnston's model is very simple and it estimates the masking threshold rather coarsely, while the masking model of PEAQ has somewhat higher computational complexity and better frequency resolution. Furthermore, the consideration of time domain masking makes PEAQ more sophisticated than Johnston's model since in the latter, this part is omitted.

This chapter concentrates on presenting the two masking models that are utilised later in this work. Even though both of these masking models are originally a part of a bigger system, i.e., a coder or a quality evaluation procedure, the examination here is mainly

restricted to the calculation of the masking threshold because that is the part needed for the speech preprocessor. The description of the models starts with Johnston's model in Section 3.1. Section 3.2 contains a rather careful mathematical presentation of the auditory masking model of PEAQ. Some other models are mentioned in Section 3.3, where the main points of the two auditory models used in the MPEG standards are also briefly viewed. Finally, Section 3.4 is a short summary of the chapter.

3.1 Johnston's model

In 1988, James Johnston announced a perceptual transform coder, named shortly as PXFM, that was designed for the efficient coding of audio signals [15]. Based on a human auditory model, the coder calculates a masking threshold and estimates the amount of quantisation noise allowed on each frequency subband so that the noise is still inaudible. The coder was one of the candidates in the standardisation performed by the ISO/IEC[1] standardisation committee during the years 1988 – 1992 that eventually produced the MPEG-1 standard [6]. By the success achieved by PXFM in the competition, the principle of Johnston's masking model ended up in the Layer III of MPEG-1 and MPEG-2 standards [28]. This can partly explain the large international interest directed at the model which has also served as motivation for this work. In this section, only those parts of the Johnston's model that are needed in the masking threshold evaluation are presented. The original application of the masking threshold in PXFM, i.e., the quantisation noise control system, is somewhat irrelevant for this work and thus it is not covered here.

PXFM is an audio coder where the masking threshold has originally been intended to be calculated for a signal sampled at 32 kHz. Frames of 2048 samples (64 ms) are extracted with an overlap of 4 ms and a square root of a Hanning window is used in the overlapping sections. The calculation of the masking threshold for each input block proceeds as follows. First, the critical band analysis of the signal is performed, by which the calculation is transferred to the Bark domain. After that, a spreading function is applied to the critical band spectrum and the spread masking threshold is calculated. The threshold is renormalised and finally combined with the absolute threshold of hearing. A block diagram of the procedure is shown in Figure 3.1 and the phases are next considered in more detail.

Critical band analysis

A fast Fourier transform (FFT) is calculated for each frame of length $N_F = 2048$ samples, yielding the complex spectrum $X(k)$, where $0 \le k \le (N_F - 1)$. The 1024 essential frequency bins of the complex spectrum are converted to the power spectrum as follows:

$$P(k) = \mathrm{Re}^2[X(k)] + \mathrm{Im}^2[X(k)],\ k = 0,..., N_F / 2 - 1. \tag{3.1}$$

[1] ISO: International Organization for Standardization
IEC: International Electrotechnical Commission

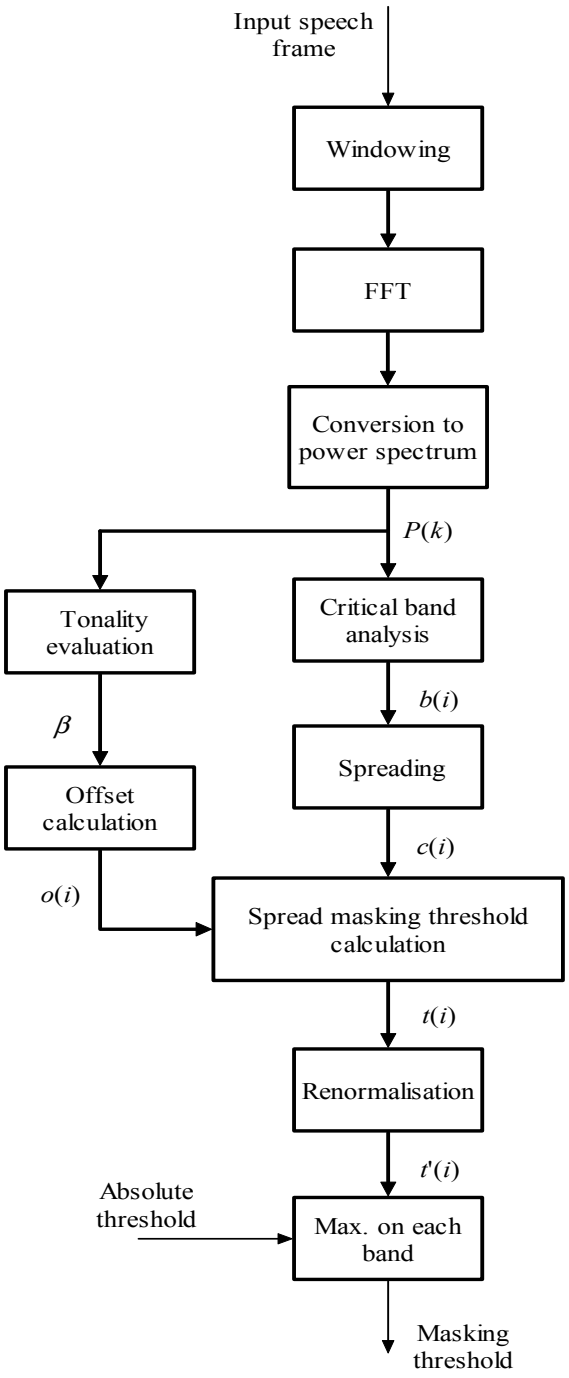

Figure 3.1. Block diagram of Johnston's masking model.

Assuming discrete non-overlapping critical bands, the energy, b, in each band is calculated by

$$b(i) = \sum_{k=\sigma_l(i)}^{\sigma_u(i)} P(k),\ i = 0,\ldots,\ i_{\max}, \tag{3.2}$$

where the lower and upper boundaries of the ith critical band, $\sigma_l(i)$ and $\sigma_u(i)$, are determined from the frequencies presented in Appendix A. The number of critical bands (i_{max}+1) depends on the sampling frequency.

Applying the spreading function

Johnston uses the spreading function proposed by Schroeder *et al.* in [33]. The spread of masking in the frequency domain is modelled by convolving the critical band energies, $b(i)$, with the spreading function, s, which has lower and upper skirts with slopes of +25 dB/Bark and –10 dB/Bark, respectively. The analytical expression for the spreading function is [33]

$$s_{\mathrm{dB}}(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2}\ \mathrm{dB}, \tag{3.3}$$

where $x = |j - i|$, i represents the Bark frequency of the masked signal and j is the Bark frequency of the masker signal. The values of the spreading function are converted from decibels to the linear scale by

$$s(x) = 10^{s_{\mathrm{dB}}(x)/10} \tag{3.4}$$

and ordered into a symmetric Toeplitz matrix, **S**, which has the form

$$\mathbf{S} = \begin{bmatrix} s(0) & s(1) & s(2) & s(3) & \cdots & & s(i_{\max}) \\ s(1) & s(0) & s(1) & & & & \vdots \\ s(2) & s(1) & \ddots & & & & \\ s(3) & & & & & & \\ \vdots & & & \ddots & & & \vdots \\ & & & & s(0) & s(1) & \\ s(i_{\max}) & \cdots & & \cdots & s(1) & s(0) & \end{bmatrix}. \tag{3.5}$$

The critical band energies are gathered into the vector $\mathbf{b} = [b(0), ..., b(i_{max})]^T$, after which the convolution can be implemented as a matrix multiplication

$$\mathbf{c} = \mathbf{S}\mathbf{b}\,, \tag{3.6}$$

where $\mathbf{c} = [c(0), ..., c(i_{max})]^T$ denotes the spread critical band spectrum.

Calculation of the spread masking threshold

In Johnston's masking model, two separate masking thresholds are used. One is for tone masking noise, estimated as $14.5 + i$ dB below the spread critical band energy, $c(i)$. The other is for noise masking a tone and it is estimated as 5.5 dB below $c(i)$ independent of the Bark frequency, i. In order to take advantage of the different cases, the tonelike or

noiselike nature of the signal needs to be determined. This can be done with the spectral flatness measure which is defined (in decibels) as

$$\lambda_{\mathrm{dB}} = 10\log_{10}\left(\frac{\sqrt[N_{\mathrm{F}}/2]{\prod_{k=0}^{N_{\mathrm{F}}/2-1} P(k)}}{\frac{2}{N_{\mathrm{F}}}\sum_{k=0}^{N_{\mathrm{F}}/2-1} P(k)}\right), \tag{3.7}$$

where the numerator and denominator are the geometric and arithmetic mean of the power spectrum, respectively. The coefficient of tonality, β, is then calculated by

$$\beta = \min\left(\frac{\lambda_{\mathrm{dB}}}{\lambda_{\mathrm{dBmin}}}, 1\right), \tag{3.8}$$

where $\lambda_{\mathrm{dBmin}} = -60$ dB. In case the λ_{dB} has a value smaller than or equal to λ_{dBmin}, the tonality coefficient is $\beta = 1$, which means that the speech segment is considered entirely tonelike. At λ_{dB} value of 0 dB, the signal is entirely noiselike and $\beta = 0$.

The tonality coefficient, β, is then used to weight the two threshold offsets mentioned above to obtain the critical band-specific offset

$$o(i) = \beta(14.5 + i) + (1 - \beta)5.5 \tag{3.9}$$

which is given in decibels. The subtraction of this offset from the spread critical band energies, $c(i)$, is performed by

$$t(i) = 10^{\log_{10}(c(i)) - o(i)/10}, \tag{3.10}$$

which gives the spread masking threshold estimate in a linear scale for each critical band.

Renormalisation of the threshold

The purpose of this step is to convert the spread threshold back to the Bark domain by undoing the convolution of the critical band spectrum, **b**, with the spreading function, s. Due to the shape of the spreading function, however, the direct deconvolution is a very unstable process which can lead to a negative energy of a threshold or other problems. Consequently, a renormalisation is used instead. Since the spreading function increases the energy estimates in each band, it can be compensated at the renormalisation stage "by multiplying each $t(i)$ by the inverse of the energy gain, assuming a uniform energy of 1 in each band [15, p. 319]". The resulting renormalised threshold is denoted by $\mathbf{t'} = [t'(0), ..., t'(i_{\max})]^{\mathrm{T}}$.

Including the absolute threshold

Finally, the masking threshold is compared with the normalised absolute threshold of hearing. The normalisation procedure utilises a tonal signal with a peak magnitude of ±1 least significant bit in a 16-bit integer and a frequency of 4 kHz. The idea is to scale the absolute threshold in such a manner that the tone is at the absolute threshold of hearing. The normalised curve is then evaluated on the Bark scale using Equation (2.3). In each critical band, the threshold, $t'(i)$, is replaced by the absolute threshold value in case the latter exceeds the former. As the absolute threshold varies inside the critical band, the mean of the values at the band edges is used. An example figure of a power spectrum and the corresponding thresholds will be presented within the implementation of the masking model in the following chapter (Figure 4.1).

3.2 Psychoacoustic model of PEAQ

The International Telecommunications Union (ITU) published in 1998 a standard for perceptual evaluation of audio quality (PEAQ) which is known as the Recommendation ITU-R BS.1387 [9]. The idea in the standard is to calculate a measure of the perceptual quality of a signal on an objective basis, utilising sophisticated models of the sensory and cognitive processes of perception. PEAQ can be used, for example, to evaluate the quality of an audio codec, and the ultimate intention has been to reduce the need for perceptual listening tests that are often laborious and expensive.

PEAQ is a rather complicated algorithm. It starts with a peripheral ear model, which is followed by several intermediate steps to be able to calculate the model output variables (MOVs). Finally, MOVs are mapped into a single value representing the quality of the test signal. PEAQ includes two different ear models; one is based on FFT and the other founds on a filter bank. Furthermore, the PEAQ model has two versions. The basic version uses only the FFT-based ear model and it has a rather high processing speed. The advanced version, using both the FFT-based and the filter bank based ear models, has higher complexity but also higher accuracy than the basic version. Due to the filter bank ear model, the advanced version has also higher temporal resolution than the basic version, but on the other hand, the FFT-based ear model gives the basic version a better frequency resolution. For the scope of this thesis, all details of PEAQ are not of interest. Rather, this section focuses on the calculation of the masking threshold as presented in the standard within the basic version of the model. In addition to the specification itself, the excellent report of the examination of the PEAQ standard written by P. Kabal [16] is referred. Small parts of it (improving the speed of the computation) have been utilised when implementing the masking model of PEAQ.

For each frame of the input signal, the calculation of the masking threshold starts with windowing, scaling and FFT. Next, the frequency response of the outer and middle ear is modelled by applying a weighting function, and the weighted spectral coefficients are then divided into critical bands. An offset is added to the critical band spectrum to account for the internal noise in the auditory system. The spread of masking (introduced in Section 2.2.4) in both frequency and time domain is taken into account by separate spreading

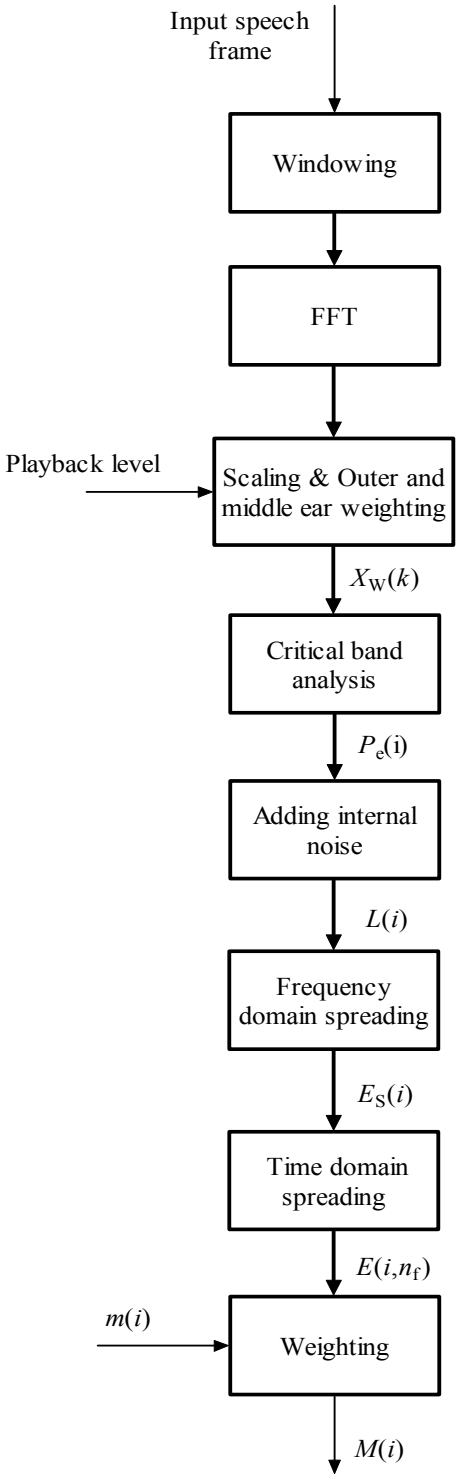

Figure 3.2. Block diagram of the masking model of PEAQ.

functions, which yields an excitation pattern. A simple weighting operation at the end is needed to calculate the final masking threshold. These steps, illustrated in Figure 3.2, will next be examined in more detail.

Discrete Fourier transform and scaling

Originally, the PEAQ model specifies the input signals to be sampled at F_S = 48 kHz. The input is cut into frames of N_F = 2048 samples (about 43 ms) with a 50 % overlap and each frame of data is windowed with a Hanning window. The frame is then transformed into the frequency domain by an FFT, yielding the signal $X(k)$ for $0 \le k \le (N_F - 1)$. To adjust the playback level of the input signal according to the assumed sound pressure level, L_p, of a full-scale test sine, the magnitude spectrum of the input frame is multiplied by a scaling factor of the form

$$g = \frac{10^{L_p/20}}{\kappa}, \tag{3.11}$$

where, according to the standard, the denominator, κ, is "calculated by taking a sine wave of 1019.5 Hz and 0 dB full scale as the input signal and calculating the maximum absolute value of the spectral coefficients over 10 frames [9, p. 36]". Kabal gives in his report a method for calculating the denominator term analytically [16]. In addition, the scaling factors needed due to the windowing and the FFT can be conveniently lumped together with the given g, yielding a combined scaling factor [16]

$$G = \frac{10^{L_p/20}}{\gamma(f_x)\frac{A_{max}}{4}(N_F - 1)}. \tag{3.12}$$

In this equation, A_{max} is the maximum amplitude of the test sine, for example 32 767 for 16-bit data, f_x = 1019.5 and $\gamma(f_x)$ = 0.8497 (for the analytical expression of $\gamma(f_x)$, see [16]). When the sound pressure level of the input signal is unknown, it is recommended to set L_p = 92 dB.

Outer and middle ear modelling

The outer and middle ear have a certain frequency response due to, among others, the meatus resonance at around 4 kHz. The response is modelled by the transfer function

$$W_{dB}(f) = -2.184(f)^{-0.8} + 6.5e^{-0.6(f-3.3)^2} - 10^{-3} f^{3.6}, \tag{3.13}$$

$$W_f(f) = 10^{W_{dB}(f)/20}, \tag{3.14}$$

where the frequency, f, is given in kHz. The shape of the outer and middle ear response curve is shown in Figure 3.3.

The weights, W, are evaluated at the same frequencies as the spectral coefficients,

$$W(k) = W_f\left(\frac{kF_S}{N_F}\right) \tag{3.15}$$

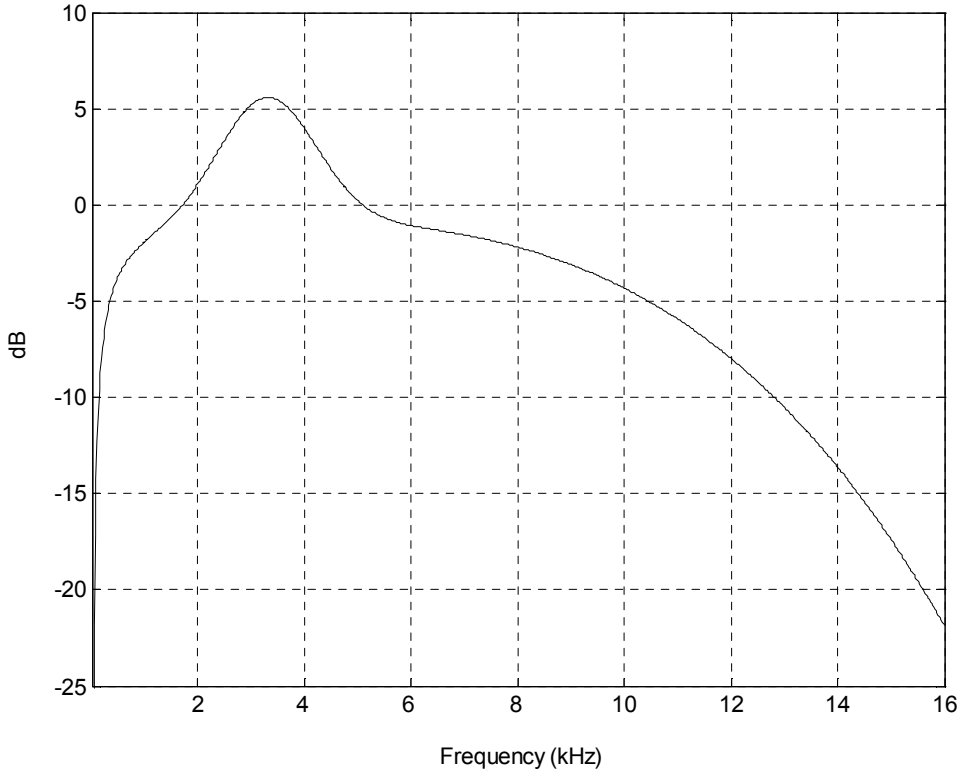

Figure 3.3. Outer and middle ear response.

and now the outer and middle ear weighted power spectrum of the input frame can be expressed by [16]

$$\left|X_W(k)\right|^2 = G^2 W^2(k)\left|X(k)\right|^2,\ 0 \le k \le N_F/2 - 1, \tag{3.16}$$

where the scaling has been taken into account in the term G.

Grouping into critical bands

Schroeder *et al.* have proposed a frequency to Bark conversion [33]

$$z = B(f) = 7\sinh^{-1}(f/650), \tag{3.17}$$

where f is expressed in hertz and z in Bark units. Utilising the given conversion and its inverse, $B^{-1}(z)$, the frequency grouping is performed in the range from 80 Hz to 18 kHz. The bandwidth is 0.25 Bark in the basic version and 0.5 Bark in the advanced version of PEAQ, resulting in 109 or 55 bands, respectively. When the lower edge of the ith frequency band is denoted by $f_l(i)$ and the upper edge by $f_u(i)$, the contribution from the kth frequency component of the FFT to the energy of the ith band is [16]

$$U(i,k) = \frac{\max\left(0,\ \min\left(f_{\mathrm{u}}(i),\ \frac{2k+1}{2}\frac{F_{\mathrm{S}}}{N_{\mathrm{F}}}\right) - \max\left(f_{\mathrm{l}}(i),\ \frac{2k-1}{2}\frac{F_{\mathrm{S}}}{N_{\mathrm{F}}}\right)\right)}{\frac{F_{\mathrm{S}}}{N_{\mathrm{F}}}}. \tag{3.18}$$

The energy in the *i*th band is

$$P_{\mathrm{e}}(i) = \sum_{k=k_{\mathrm{l}}(i)}^{k_{\mathrm{u}}(i)} U(i,k)\left|X_{\mathrm{W}}(k)\right|^2, \tag{3.19}$$

where $U(i,k)$ is non-zero in the interval $k_{\mathrm{l}}(i) \le k \le k_{\mathrm{u}}(i)$, i.e. between the lower and upper boundaries of the *i*th band. The energy is finally limited to a value greater than or equal to $1 \cdot 10^{-12}$.

Adding internal noise

To model the internal noise generated in the ear, a frequency dependent offset is added to the energies of the frequency groups. This process is incorporated in the following equation by which the pitch patterns are obtained in dB SPL:

$$L(i) = 10\log_{10}\left(P_{\mathrm{e}}(i) + 10^{0.1456(f_{\mathrm{c}}(i)/1000)^{-0.8}}\right). \tag{3.20}$$

The frequency units are hertz and $f_{\mathrm{c}}(i)$ (in hertz) denotes the centre frequency of the *i*th band. As remarked in [36], the internal noise in the equation above, together with the outer and middle ear frequency response modelled by the equation (3.13), approximate the threshold in quiet as presented by Terhardt in [35].

Frequency domain spreading

The spread of energy in the frequency domain is modelled by applying to the pitch patterns a spreading function adopted from [35]. When presented in decibels, the spreading function is triangular, with the peak of the triangle at the centre frequency of the target band. The lower slope is constant but the upper slope is energy and frequency dependent:

$$S_{\mathrm{dBl}}(i,j) = 27(i-j)\Delta z, \quad i \le j, \tag{3.21}$$

$$S_{\mathrm{dBu}}(i,j,L) = \left(-24 - \frac{230}{f_{\mathrm{c}}(j)} + 0.2L(j)\right)(i-j)\Delta z, \quad i \ge j, \tag{3.22}$$

where energy is spreading from the bands *j* to the band *i* and Δz has the value 0.25 in the basic version and 0.5 in the advanced version.

In the calculation of the spreading effect, a few auxiliary functions are needed. First, since the spreading function increases the energy estimates in each band, the energy gain is normalised to unity by

$$A(j,L)=\sum_{i=0}^{j-1}10^{S_{\mathrm{dBl}}(i,j)/10}+\sum_{i=j}^{N_{\mathrm{C}}-1}10^{S_{\mathrm{dBu}}(i,j,L)/10}\,, \tag{3.23}$$

where N_{C} denotes the number of critical bands (109 or 55 depending on the version). For notational convenience, the symbol $L(j)$ that appears in Equation (3.22) was replaced by mere L in the equation above, and this continues also in the following two equations. The spreading function can now be expressed as [16]

$$S(i,j,L)=\begin{cases}\dfrac{1}{A(j,L)}10^{S_{\mathrm{dBl}}(i,j)/10}, & i\le j,\\ \dfrac{1}{A(j,L)}10^{S_{\mathrm{dBu}}(i,j,L)/10}, & i\ge j.\end{cases} \tag{3.24}$$

Another normalising factor is calculated with a reference level of 0 dB for each band [16]:

$$B_{\mathrm{S}}(i)=\left(\sum_{j=0}^{N_{\mathrm{C}}-1}\left(S(i,j,L_0)\right)^{0.4}\right)^{1/0.4}\,, \tag{3.25}$$

where $L_0 = 0$ dB, which means that all frequency bands are assumed to have unit energy. Finally, the spread energy in band i is obtained by [16]

$$E_{\mathrm{S}}(i)=\frac{1}{B_{\mathrm{S}}(i)}\left(\sum_{j=0}^{N_{\mathrm{C}}-1}\left(L(j)S(i,j,L(j))\right)^{0.4}\right)^{1/0.4}\,. \tag{3.26}$$

This Bark domain energy response is referred to in the standard as “unsmeared excitation patterns [9, p. 45] ”.

Time domain smearing

Backward masking is not taken into account in the basic version of PEAQ but forward masking is simulated by smearing the energies in each frequency group by a first order low-pass filter. The time constant of the filter used for the ith frequency band is computed as

$$\tau(i)=\tau_{\min}+\frac{100}{f_{\mathrm{c}}(i)}\left(\tau_{100}-\tau_{\min}\right), \tag{3.27}$$

where $\tau_{100} = 0.030$ s and $\tau_{\min} = 0.008$ s. The former represents the time constant at 100 Hz and the latter approximates the time constant at the highest centre frequency (i.e., at about

17.69 kHz). Since the time domain spreading depends on the previous frame in addition to the current frame, a frame index, n_f, is introduced in the representation of the unsmeared excitation patterns, $E_S(i, n_f)$. The excitation patterns are calculated by

$$E_f(i,n_f) = \alpha(i)E_f(i,n_f - 1) + (1-\alpha(i))E_S(i,n_f), \tag{3.28}$$

where $\alpha(i)$ is calculated from the time constant $\tau(i)$ by [16]

$$\alpha(i) = e^{-1/(F_{ss}\tau(i))}, \tag{3.29}$$

and the frame rate is [16]

$$F_{ss} = \frac{F_S}{N_F/2}. \tag{3.30}$$

The initial condition for the filtering can be stated as $E_f(i,-1) = 0$ for all i. To improve the temporal resolution, the filtered values of the excitation patterns, $E_f(i, n_f)$, are replaced by the corresponding unfiltered values, $E_S(i, n_f)$, in case the latter exceeds the former. The excitation patterns are then

$$E(i,n_f) = \max\left(E_f(i,n_f), E_S(i,n_f)\right). \tag{3.31}$$

Masking threshold

Finally, the masking threshold is calculated by weighting the excitation patterns with

$$m(i) = \begin{cases} 3.0, & i\Delta z \le 12, \\ 0.25\, i\,\Delta z, & i\Delta z > 12, \end{cases} \tag{3.32}$$

obtaining the masking threshold:

$$M(i) = \frac{E(i)}{10^{m(i)/10}} \tag{3.33}$$

which is referred to in the standard as “mask patterns [9, p. 45]”.

3.3 Other models

In addition to the two psychoacoustic models presented above, many other models have naturally been published. For example, a model developed by Soulodre has been presented in [37]. It provides a masking threshold as a part of a system that has been designed to remove camera noise from film soundtracks. A method for measuring perceptual audio quality differently from the PEAQ recommendation of ITU has been published in [2]. It

contains an auditory model that is applied to evaluate the internal (i.e., psychophysical) representation of the signal. Yet another model, developed for the purpose of noise shaping in audio coders, can be found in [3]. However, particularly well-known auditory models are those used in the Moving Picture Experts Group (MPEG) standards, and they will be discussed a little further here. Basically, two different ideas can be found in the MPEG standards, both introduced in MPEG-1 [6], where they are denoted psychoacoustic models 1 and 2. The Advanced Audio Coding (AAC) standard in MPEG-2 [7] uses an auditory model derived from the second model of MPEG-1 and furthermore, MPEG-4 [8] uses AAC. An overview of the two different MPEG models and their usage is provided in the following.

MPEG-1 contains three layers (I – III) such that the higher layers utilise parts of the lower ones. Increased complexity and delay are encountered when moving to the higher layers but on the other hand, considerable bit savings are also achieved. The supported audio sampling rates are 32 kHz, 44.1 kHz and 48 kHz. Audio signals are coded on a subband basis, and the psychoacoustic model 1 determines, for each subband, the level of the signal and the minimum masking threshold. The output of the model is the signal-to-mask ratio (SMR) that is utilised in the bit or noise allocation of the subbands. The procedure starts with Hanning windowing and the calculation of the power spectrum from which the sound pressure level in each subband is computed. For the determination of the masking threshold, the tonal components of the spectrum are found by seeking the local maxima. The tonal and non-tonal components of the spectrum are discriminated, and from both of these groups, only the salient masking components are saved in a decimation procedure. Then, the individual masking threshold of each tonal and non-tonal masker is calculated by summing the sound pressure level of the masker with a negative offset and the spreading function contribution. In the global masking threshold, these separate thresholds and the absolute threshold of hearing are combined and finally, the minimum masking threshold and SMR in each subband are determined. Throughout the description of the psychoacoustic model 1 [6], slightly different versions for Layer I and the more sophisticated Layer II are provided. The model can also be adapted to Layer III.

The second psychoacoustic model of MPEG-1 does not extract tonal and non-tonal masker components from the spectrum, but rather it resembles the auditory model proposed by Johnston [15]. It partitions the energy of the spectrum into perceptually relevant bands, estimates the tonality inside the input block and incorporates convolution to account for the spread of masking. However, in the model of MPEG-1, the tonality factor is based on the predictability of the spectral coefficients and it is computed for each band in the threshold calculation partition domain. The frequency resolution in the partition domain is either one FFT line or 1/3 critical band, whichever is wider, resulting in up to 59 bands when the sampling frequency is 48 kHz. In Layer III, the psychoacoustic model includes also pre-echo control and some additional modifications due to the switching between long and short block types. The long blocks are used during the stationary parts of the signal while the short block type is used to handle the transients. The psychoacoustic model 2 is mainly used in the Layer III while the lower layers usually use the previous model.

The psychoacoustic model of AAC has been derived from MPEG-1 model 2 with very subtle modifications. The AAC psychoacoustic model handles attacks in the same manner

as the model 2 within Layer III; switching from long to short block type is based on the comparison between the perceptual entropy measure and a predetermined switching threshold. However, the length of the blocks is different in AAC and MPEG-1 Layer III. Furthermore, AAC provides progressive scalability by supporting up to 12 different sampling frequencies ranging from 8 kHz to 96 kHz. Slight differences between these models are also found in the calculation of the unpredictability measure in the upper part of the spectrum. The outputs of the perceptual model of AAC include, among others, the SMR values, energy thresholds, block type information and an estimation of the number of bits needed to encode the current block of samples. The AAC syntax is fully supported by MPEG-4 General Audio coding and, specifically, the psychoacoustic model of AAC is used in MPEG-4.

3.4 Summary

This chapter presented some psychoacoustic models that are widely used in speech and audio coding applications. The two most salient models for this work are the model proposed by Johnston and that used in PEAQ and, therefore, the mathematical contents of these models were described rather carefully. To complement the concept of the present state of the auditory modelling within signal processing, the psychoacoustic models of the MPEG standards were also viewed on a coarse level and some other models were mentioned very briefly. The concern in the following will, however, be on the Johnston's model and PEAQ. The application of the masking models within this thesis will be described in the following chapter.

4 Implementing preprocessor

Preprocessing of speech or audio signals typically aims at facilitating the coding process before initiating the actual coding. This objective is pursued in this work by taking advantage of the properties of the human auditory system. The idea is to make the speech signal applicable for more efficient coding by removing those elements of the signal that cannot be perceived by a human listener.

Traditionally, preprocessing is needed if the signal to be coded has been disturbed by background noise or interfering sounds. Especially low-bit-rate speech codecs that utilise models of human speech production are sensitive to the deviations of the input signal from pure speech. One approach to alleviate this problem by preprocessing has been proposed in [17]. It modifies the input signal according to an objective criterion, striving for an acceptable compromise between the quantisation error and the distortion caused by the modification. Acoustic noise suppression techniques that exploit auditory models have also been widely examined, e.g. [37, 38, 40, 42]. Furthermore, performance improvement in perceptual audio coding has been reported in [4], where a psychoacoustic pre- and postfilter for a transform coding scheme is introduced. The preprocessor proposed in this thesis does not, however, aim at suppressing background noise. Rather, it assumes clean input speech and its purpose is to enable the reduction of the bit rate in speech codecs—without degrading the speech quality—based on the fact that the irrelevant information does not need to be coded. Alternatively, improvements in the output quality of speech codecs can be sought without increasing the bit rate.

The main objective of this chapter is to introduce an implementation that performs the perceptual irrelevancy removal. All the functions have been implemented in Matlab. The procedure starts with the selection of the masking model which is considered in Section 4.1. In Section 4.2, some useful modifications to the existing model are proposed. Using the resulting model, the masking threshold is calculated in order to determine the simultaneously masked components of the input spectrum. These perceptually irrelevant components are removed by adaptive filtering and the modified speech is finally converted into time domain. Section 4.3 provides a detailed description of the removal of the masked frequencies and takes a view on the overlap-add method that is used as the basis when

processing successive speech frames. In Section 4.4, the computational load and the delay of the implementation are discussed. Section 4.5 considers some other possible applications for the masking threshold that could be tested for the purpose of irrelevancy removal. Brief conclusions of this chapter are drawn in Section 4.6.

4.1 Choosing of model

The speech preprocessing block implemented in this work was practically built around an auditory model. From the two widespread models presented in detail in Chapter 3, one was chosen for the implementation of the speech preprocessor. Being such an essential part of the preprocessor, the selection of the model inevitably affects many of the properties of the final implementation. Depending on the operating environment, the computational complexity can be a limitation and, indeed, the differences in the computational load between different auditory models can be rather large. Furthermore, it is desirable that the masking model does not bring about additional delay to the processing. The resolution in the Bark domain varies between models, which implies that the accuracy at which the masking threshold is accommodated to the spectrum of the input segment is different in separate models. Naturally, the overall modelling of the human auditory system is done with variable accuracy; for example, the time domain masking is totally omitted in Johnston's model. Thus, the auditory models can be considered from several different angles and the choice is finally made according to the particular situation, often ending up in a compromise.

In this work, the auditory models under consideration were basically the Johnston's model and the PEAQ model. They were both implemented with Matlab, whereas the other options were left out. For example, implementing the model used in the MPEG-4 standard was not considered reasonable because of its complexity and because it would also require some additional delay to choose between the short and long input block type. From the two implemented masking models, Johnston's model was found to be better eligible for the speech preprocessing application of this work. In the following, some underlying facts are discussed.

The speech quality produced by the two different implementations was examined by a great deal of informal listening done by the author. It was found that the speech quality obtained with the preprocessor that was based on the PEAQ model was considerably worse than expected. A few other experienced listeners also confirmed this. The masking threshold generated by this preprocessing function was very high at the upper part of the spectrum, which seriously deteriorated the speech quality. Tests with different speech samples indicated that even 80 % of the spectral coefficients per frame were deemed masked in this system. Corrections to this problem were sought but they would have taken the model far away from the original and, consequently, the carefully designed functions modelling the human perception would no longer have been completely valid. Since compromising the accuracy of the perceptual modelling was considered disadvantageous in this application, the development was aborted. Further research to alleviate the problems was left to the future considerations. In conclusion, the masking model of PEAQ that was originally designed for audio signals seemed not to be readily applicable to

narrowband speech. The quality degradation was unacceptable since it was contrary to the original idea of the nearly transparent quality at the output of the preprocessor. Based mainly on the speech quality aspect, the masking model proposed by Johnston was finally chosen as the basis of the preprocessor.

4.2 Modifications to masking model

Originally, Johnston's application assumed that the input signal was sampled at 32 kHz and used a data window length of 2048 samples (64 ms) with an overlap of 1/16th. In the narrowband compatible implementation, i.e., using a sampling frequency of 8 kHz, a frame length of 320 samples (40 ms) and an overlap of 50 % between the frames are used. Each input frame is weighted by a Hamming window as proposed in [21] and transformed into the frequency domain representation via the discrete Fourier transform (DFT). As the length of DFT should be a power of two to be able to use the more efficient fast Fourier transform (FFT), the length of 512 samples was chosen. It provides a frequency resolution of 15.625 Hz which, compared to the smallest critical bandwidth of 100 Hz, is well adequate for this application. With the highest frequency in the input being limited to 4 kHz, the number of critical bands is 18.

The calculation of the offset in the implementation does not exactly follow Equation (3.9) presented in Section 3.1. This is partly due to the finding published in [20] that the masking threshold, when calculated using Equation (3.9), tends to be too high in frames that have a very flat spectrum. Therefore, the whole frame, or at least too large a part of it, can be deemed masked, which deteriorates the output quality. To overcome this problem, the following modification to the offset (in decibels) has been suggested [20]:

$$\begin{aligned} o(i) &= 100, && \beta < 0.2, \\ o(i) &= (1.1 - \beta)(15 + i/2), && \beta \geq 0.2. \end{aligned} \tag{4.1}$$

In the equation above, i denotes again the critical band number. By the presented equation, the frames that contain noiselike signal, i.e., that have a small coefficient of tonality, β, will have a very low masking threshold due to the large offset. Hence, the spectral coefficients in these frames are spared from being masked.

During the work within this thesis, however, even more modifications of the offset were found to be necessary. That is because the coefficient of tonality in the realized implementation is very often equal to unity, which produces very small offsets when Equation (4.1) is used and, consequently, many frames of the input speech are deemed totally masked. This appears as silence in the output since the preprocessor removes the masked frequency bins. To avoid this, the offset is returned to the original form (Equation (3.9)) in case $\beta = 1$, resulting in $o(i) = 14.5 + i$ dB in those frames. The rest of the masking threshold determination follows the procedure detailed in Section 3.1. An example of the thresholds for a female speech frame is shown in Figure 4.1.

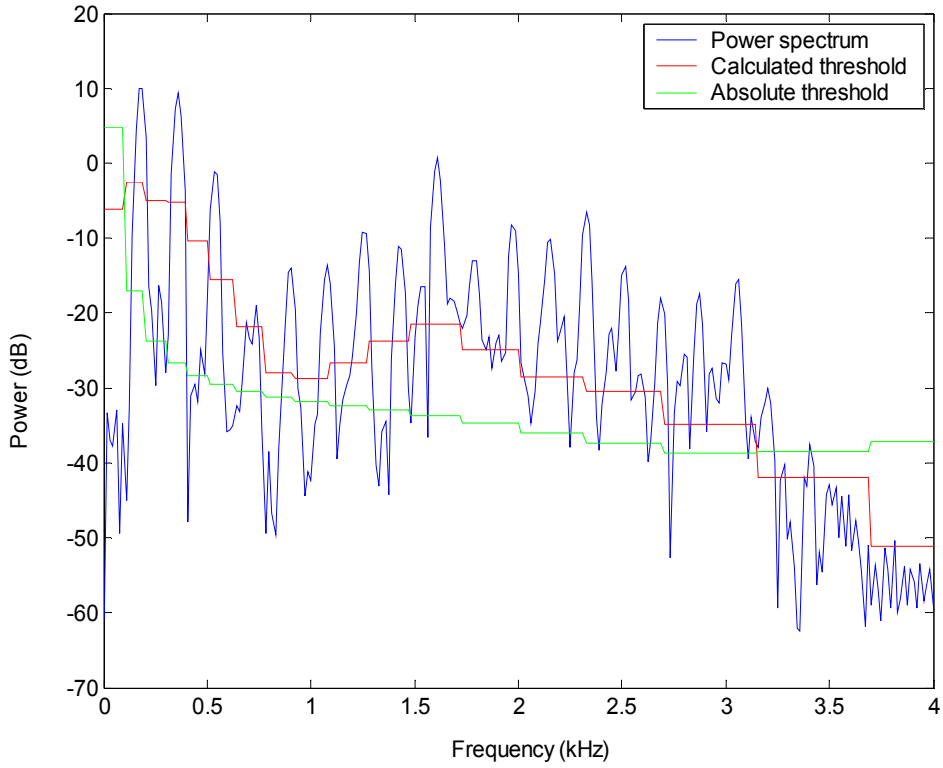

Figure 4.1. Example of the masking threshold determination.

4.3 Removal of masked components

After the masking threshold has been calculated for the input frame, it is compared with the power spectrum components in the corresponding frame to produce a binary mask. The mask is set to a value of zero at those frequencies where the power spectrum is below the masking threshold and a value of one is used elsewhere. Figure 4.2 (a) presents an example of the mask. A straightforward means to remove the masked frequencies would then be the multiplication of the complex spectrum of the frame by the mask at each frequency. This corresponds to adaptive filtering of the input speech since the filter frequency response, i.e., the mask, changes from frame to frame. In general, the adaptation of the filter frequency response easily causes time domain aliasing to the output when filtering non-stationary signals, because the impulse response of the filter becomes too long [32]. Time domain aliasing results in the output signal to be deteriorated by non-harmonic distortion that spreads over the whole frequency range [32]. Also considering the fact that the adaptation of the filter in the application presented here comprises rather violent changes in the filter frequency response due to the binary character of the mask, an alternative solution to the spectrum modification should be sought.

To reduce the time domain aliasing, a method proposed by Rass and Steeger is applied. As mentioned above, the main problem is the length of the filter impulse response, L, that should fulfil the condition

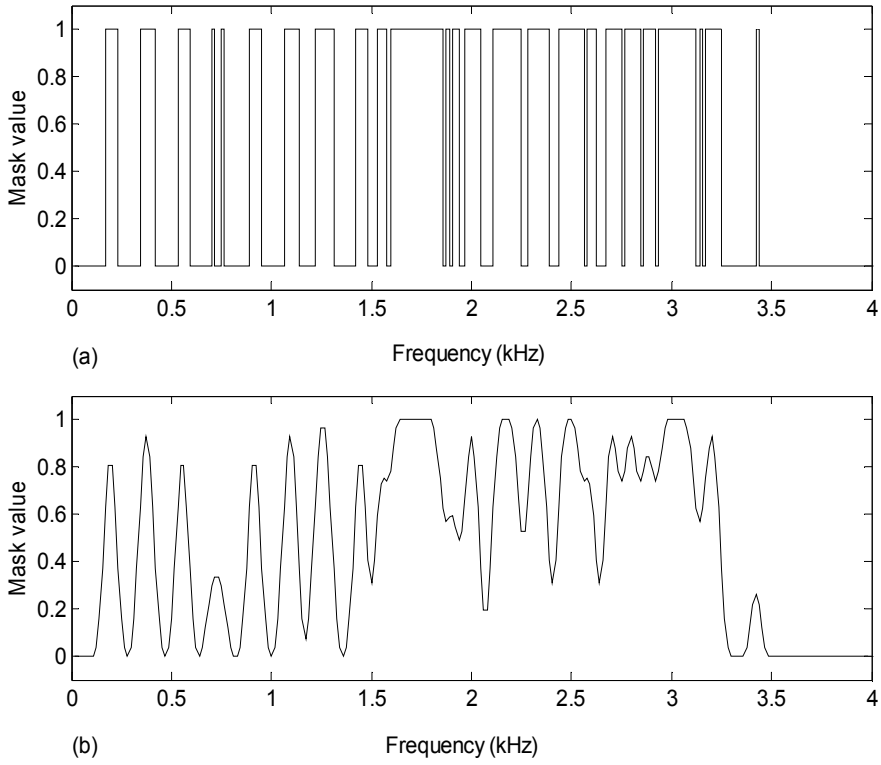

Figure 4.2. Binary mask (a) and its smoothed version (b), corresponding to the speech frame shown in Figure 4.1.

$$L \leq M - N + 1 \tag{4.2}$$

in order to avoid the time domain aliasing. In the equation, N denotes the length of the input data segment that is converted to the frequency domain by a DFT of length M. The simplest way to reduce the aliasing problem would obviously be the use of large M and small N, but that is computationally ineffective. Another solution involves the transformation of the filter frequency response into the impulse response by the IDFT and shortening this response to the desired length by windowing. Finally, the input segment could be filtered in the time domain by convolution. Unfortunately, this technique requires very much computation, which makes it less attractive in practical applications. [32]

The final solution proposed in [32] is the convolution of the filter frequency response with an appropriate window. In this method, the computational load can be traded off against the attenuation of the aliasing components, yielding a good compromise. To make the convolution less demanding to compute, the window should have only a small number of non-zero components in the frequency domain. The time domain representation of the window is therefore not limited in length, but it has a mainlobe and sidelobes. To simulate the truncation of the filter impulse response with the highest possible accuracy, the window should be chosen such that most of its energy is concentrated in the mainlobe and the sidelobes are small. The digital prolate spheroidal sequence (DPSS) [41] is an optimal choice in this application, leading to a maximised ratio of the desired signal power to the distortion signal power. [32]

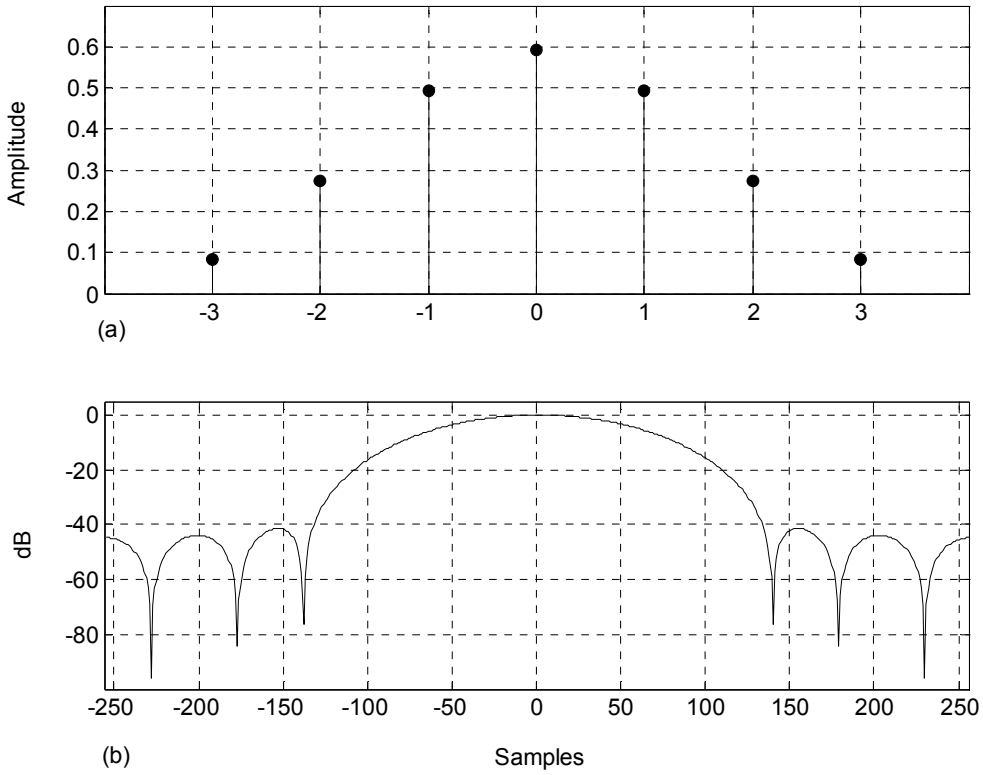

Figure 4.3. Chosen DPSS in frequency domain (a) and in time domain (b).

The design procedure of DPSS, detailed in [41], has been implemented in this work. Basically, DPSS is determined as the eigenvector that corresponds to the largest eigenvalue of a certain symmetric matrix (see [41]). The contents of the matrix depend on the two design parameters, namely the window length and the relative peak sidelobe height. These parameters control the mainlobe width. For example, it can be reduced by mitigating the sidelobe attenuation and by increasing the length of the window. In this work, the window length was set to 7 samples and the relative sidelobe height to –39 dB. This selection is quite prudent as the aim is to avoid excessive smearing of the mask. Figure 4.3 (a) presents the chosen window in the frequency domain, i.e., in the form that is used in the convolution. The time domain equivalent, shown in Figure 4.3 (b), has been calculated by the IFFT of the zero-padded DPSS and normalised to the maximum of 0 dB. By Equation (4.2), it is desirable to obtain a fair amount of attenuation outside the range of width 512 – 320 + 1 = 193 samples. It can be seen from Figure 4.3 (b) that the attenuation at the edges of the segment of this width is about 15 dB. The smoothing is finally done by computing the circular convolution between the mask and DPSS. As a result, the frequency response of the adaptive filter is no longer a binary sequence but a smoother version of it. Figure 4.2 gives an example of the mask before and after the smoothing.

The next operation is the removal of the frequency components that have been deemed perceptually irrelevant due to the masking. It is done by multiplying the complex spectrum of the frame by the smoothed mask at each frequency. As the last phase of the actual preprocessing, the output segment is transformed into the time domain using IFFT. The

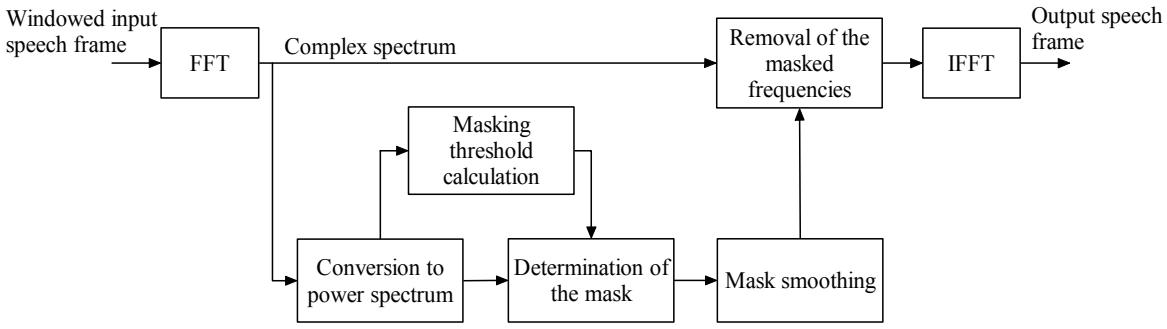

Figure 4.4. Block diagram of the method of preprocessing one frame of speech.

main functional parts of the perceptual irrelevancy removal method are summarised in the form of a block diagram in Figure 4.4.

The overlapping sections are combined by adding together the latter half of the previous speech segment and the first half of the current segment, as illustrated in Figure 4.5. The overlapped segment (from x to $x + 160$ in the figure) is available for the output as soon as it has been normalized so that the effect of the windowing is cancelled. This is done by dividing the new output segment sample-wise by a vector that has been determined beforehand as the sum of the lower and upper slopes of the Hamming window. The latest 160 samples of the output signal remain half-finished waiting for a new frame to be processed.

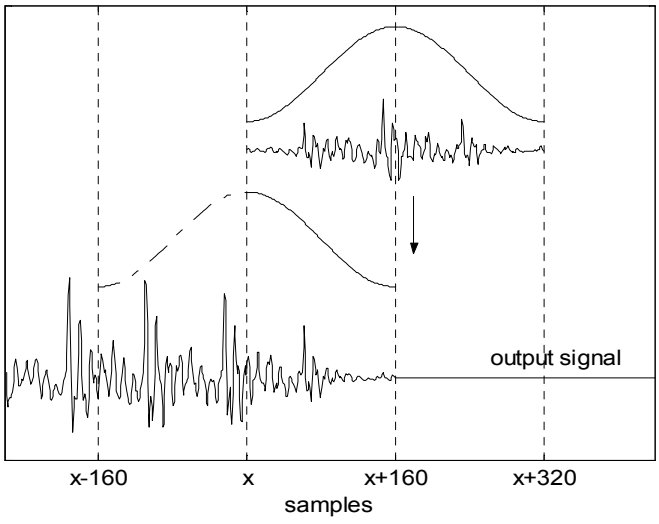

Figure 4.5. Construction of the output signal by overlap-add. The latest output segment is added to the end of the half-finished output.

4.4 Analysis of preprocessor

When considering speech or audio codecs, the four most important dimensions of performance are usually signal quality, bit rate, communication delay and the computational complexity of the algorithm [14]. The same factors are next used to evaluate the performance of the proposed preprocessor. However, bit rate cannot be assessed here because the preprocessor does not involve any kind of signal coding. The possible bit savings achieved by the usage of the preprocessor in the front end of a speech codec will be approximated in Chapter 5. In addition, the speech quality aspect will be treated in Chapter 5 in connection with the other test results.

The preprocessing functions presented in this work were implemented in Matlab, but they were not exactly optimised for the minimal computational load. However, the computational complexity was evaluated to give some indicative parameters of the proposed preprocessing method and its possible alternative that would use the masking model of PEAQ. The computational load was assessed by estimating how many floating-point operations are needed for producing one second of output speech (sampled at 8 kHz). The following approximations were used: Trigonometric and logarithmic operations, as well as the raising of a number to a specified power, require eight floating-point operations each. Square root and division require four and other arithmetic operations one floating-point operation.

The implementation using Johnston's model seemed not to possess a great computational load. It was calculated by hand from the source code that about 3.38 million floating-point operations are required to preprocess 8000 samples of speech. If the Johnston's model were switched to the masking model of PEAQ, the computational load would increase to about 3.94 million floating point operations per second. In both cases, the FFT and its inverse are the most complicated parts of the processing. It should be noted that the given figures are rather approximative and they do not include the initialisation operations. By optimising some parts of the implementation, the complexity could be slightly reduced. An implementation in C language is worth considering in the future development.

Another feature of the preprocessor that can be presented objectively is the delay. Due to the overlap, the lookahead needed by the preprocessing function is 20 ms. The algorithmic delay is 40 ms. Using the same overlap-add procedure, the lookahead and the delay remain the same if the psychoacoustic model is taken from PEAQ. Since the frame length, 40 ms, is rather long compared to the frame lengths traditionally used in speech coding applications, the possibility to use shorter frames could be worth studying. The frequency resolution would inevitably deteriorate, but it should be examined whether this has a significant impact on the operation of the preprocessor. Different amounts of overlap should be tested within the same experiment. It is rather likely that some reduction of the algorithmic delay could be attained by reasonable effort.

The average number of masked components in each frame was examined for a speech signal of which 39 minutes were spoken by female and 31 minutes by male speakers, altogether 10 persons. Table 4.1 presents the results in percentages and the weighted average values, the weights being the relative durations of the segments in the total signal.

Table 4.1. Percentages of the masked components in 70 minutes of Finnish speech.

	1	2	3	4	5	Average	Combined average
Female	42.4	49.9	52.1	44.5	50.6	47.0	43.3
Male	37.3	38.1	41.2	38.7	37.3	38.0	

It can be observed from Table 4.1 that consistently more frequency components are deemed masked in the female than in the male speech. However, this does not cause considerable differences in the perceptual quality of preprocessed male and female speech. The speech quality will be further examined in Chapter 5.

4.5 Other possible applications for threshold

Many different applications for masking models have been proposed since the beginning of their emergence and development in the late seventies. This section briefly presents two applications that, from the viewpoint of the work done for this thesis, might possess some interesting future research potential. First, a method that uses the masking threshold in the removal of perceptual irrelevancy within linear prediction coding (LPC) is presented. Second, the possibility to utilise the threshold in the quantisation of spectral components is considered.

In contrast to the idea of completely removing the masked components from the speech signal, a milder version has been proposed by Lukasiak and Burnett [19 – 22]. They have tested a method in which perceptually irrelevant components of the speech signal are omitted during the LPC coefficient calculation but the original (i.e., non-modified) speech is used in other parts of the encoding. The method utilises simultaneous masking by incorporating the masking model proposed by Johnston and it has been reported to yield improvements in the perceptual quality of coded speech without increasing the bit rate. The improvement is based on the fact that the perceptually relevant unmasked parts of the LPC spectrum are better modelled by the modified than by the standard LPC coefficients. Consequently, the new LPC residual contains less perceptually important information, and the speech signal can be coded more accurately than the original signal with the same bit rate. Since the technique described above lies on the same kind of irrelevancy removal method as utilised in this thesis, there might exist a basis for interesting future research. The work presented here could be continued by combining the modified LPC analysis with a standardised speech codec, for example the code-excited linear prediction (CELP) codec presented in [5], and experiments could be conducted to determine how many bits per second can possibly be saved by this method.

Another possible future research direction might contain examination on how the masking threshold can be utilised in spectrum quantisation, for example, within the waveform interpolation (WI) coder [18]. As mentioned in the previous chapter, one of the most common applications of masking models has been the controlling and shaping of quantisation noise in speech and audio coders in such a manner that human listeners could not perceive it. When applied to the WI coder, the masking threshold could be used to assist the quantisation of the slowly evolving waveform (SEW) that represents the voiced component of the LPC residual [27]. At low bit rates, it is useful to present SEW by the

phase and magnitude spectra and to transmit only the magnitude [18]. In order to optimise the speech quality, it would be advantageous to evaluate the perceptual relevance of the different magnitude components by comparing them with the masking threshold and to use this information when searching for the best code vectors for SEW.

4.6 Summary

In this chapter, the implementation of a perceptual preprocessing method for narrowband speech signals was presented. The purpose of the preprocessor is to remove perceptual irrelevancy from speech signals to enable more efficient coding. After discussing the properties of the masking models described in Chapter 3, the Johnston's model was chosen as the basis for the preprocessor. The necessary modifications to the original model were proposed and the implementation of the irrelevancy removal method was carefully described. The computational complexity of the preprocessor was assessed as one of the factors affecting the performance of the system. Finally, some future research possibilities within the masking threshold were outlined. The ability of the preprocessor to achieve its objectives will be examined in the following two chapters.

5 Experiments and results

The previous chapters have described the implementation of the preprocessing technique that aims at removing the perceptually irrelevant frequency components of the speech signal. The psychoacoustic model, which comprises the basis of this method by providing the masking threshold, has been detailed in Chapter 3 and the actual implementation in Chapter 4. The purpose of the preprocessor is to modify the speech signal in such a manner that it could be coded more efficiently than the original signal. This twofold objective can refer either to bit rate reduction without perceivable speech quality deterioration or to better speech quality without increasing the bit rate. Decrements in the coding delay or in the computational complexity can also be regarded as improvements in the coding efficiency, but they are not considered here. It is obvious that an additional processing stage in front of a codec increases the total computational burden and possibly also the delay.

In order to find out how well the implemented preprocessor meets the objective, several tests were performed. First, the performance of the preprocessor itself, with no speech codec attached to it, was examined through objective measurements and perceptual evaluation. The aim was to study the compressibility and the perceptual quality of the preprocessed speech signal. Second, the preprocessed speech was used as an input signal to a speech codec and another set of objective measures were calculated. The second part also contained a listening test to reveal the possible end users' opinion about the quality of the speech that has first been preprocessed and then fed through a codec. The following two sections present the test procedures and the results, first Section 5.1 without a speech codec and then Section 5.2 with two different codecs. Section 5.3 concludes this chapter.

5.1 Evaluation of preprocessor

Preliminary listening and adjustment of the preprocessor were performed until the speech quality at the output of the preprocessor was, for the most part, equal to that at the input. The main modifications that were found necessary were those of the masking model; they have been described in Section 4.2. Then, the effect of the perceptual preprocessing on the

compressibility of the speech signal was studied through some objective measurements. Due to the key position of the linear predictive coding (LPC) in almost all modern day speech codecs, the LPC analysis was applied also in this experiment. All the details about the linear prediction (LP) will not be studied here but an excellent description can be found, for example, in [30]. The LPC analysis was performed on both the original and the preprocessed speech and the resulting residuals were examined as described in Subsection 5.1.1. To finally verify that the preprocessing does not cause substantial deterioration of the speech quality, an informal listening test was performed using the comparison category rating (CCR) procedure [11]. The settings of the preprocessor were maintained as they were during the objective measurements in which an average of 43 % of the spectral components were deemed masked in each frame. Subsection 5.1.2 presents the CCR procedure and the results of the perceptual evaluation.

5.1.1 Objective measurements

The coefficients of the 10th order linear prediction filter were estimated for every 20 ms speech segment by the autocorrelation method. A 30 ms asymmetric window, composed of Hamming and cosine windows, was used. The autocorrelations were multiplied by a lag-window containing white noise correction and a bandwidth expansion of 60 Hz before calculating the filter coefficients using the Levinson-Durbin recursion (see, for example, [31]). The 10th order LP polynomial was only used to form the LPC analysis filter by which the residuals were obtained from the speech signals. The coefficients of the filter were excluded from the actual tests because it has already been shown in [21] that the LPC coefficients calculated from the modified speech signal have quantisation properties very similar to those of the standard LPC coefficients. In other words, no extra bits are required in their quantisation compared to the LPC coefficients calculated from the original speech [21].

A Finnish speech signal, consisting of about 39 minutes of female and 31 minutes of male speech, was preprocessed using the method described in Chapter 4. The LPC residual signals of the original and the preprocessed speech were produced and the energies of the different residuals were computed by

$$e(r) = \frac{1}{N} \sum_{n=0}^{N-1} r(n)^2 \,, \tag{5.1}$$

where r denotes the residual signal consisting of N samples. The energy of the residual calculated from the preprocessed speech was, on average, 20.3 % smaller than that of the residual obtained by analysing the original speech. The decrease in the energy causes modest loudness differences between the original and the preprocessed speech, but this effect can be compensated at the decoder using an additional level control. By utilising an adaptive control system, the level adjustment does not increase the amount of data to be quantised. This kind of a control system has not been implemented in this work but it might be a part of the future development.

Another measure that was evaluated was the entropy of each residual. The entropy is calculated from the probabilities, p, of the different code words of a signal in its m-bit presentation by

$$H = -\sum_{i=0}^{M-1} p(i)\log_2 p(i), \tag{5.2}$$

where $M = 2^m$. By using the preprocessed instead of the original speech in the LPC analysis and filtering, the entropy of the 16-bit residual signal was reduced by 7.2 %. The entropy of a signal indicates the lower limit of the number of bits necessary for the lossless coding of the signal. In this case, the absolute reduction of the entropy was about 0.63 bits per sample, which would provide a saving of 5.0 kb/s if lossless coding was applied to the residual. This result together with the energy reduction presented above confirms that the usage of the preprocessing technique in the front end of a speech coder produces a residual signal that can most likely be compressed more efficiently than the original residual signal.

5.1.2 CCR listening test

Throughout the objective measurements described above, the preprocessing function was kept at such constant settings that the output speech did not suffer from substantial deterioration. This was verified through a CCR listening test [11] which was conducted in co-ordination with M.Sc. Ulpu Sinervo at Tampere University of Technology. In the CCR test, listeners are presented with pairs of speech samples and, for each pair, they are asked to grade the quality of the latter sample with respect to the former on a seven-point scale. The grades of the scale together with their explanations are shown in Table 5.1. Each pair contains a processed sample and a quality reference that are presented in random order.

Table 5.1. Grades of the CCR test.

Grade	Meaning
3	much better
2	better
1	slightly better
0	about the same
– 1	slightly worse
– 2	worse
– 3	much worse

In this experiment, each sample was a single sentence. The test material consisted of Finnish speech filtered with an intermediate reference system (IRS) filter. Six female and six male speakers were chosen from a database in which the sampling frequency was 16 kHz. The samples for the listening test were generated by taking one sentence from each speaker and processing the signals as follows. First, the level of each speech sample was adjusted using software based on the ITU-T Recommendation P.56 [10] (described further in [13]). Then, the sampling frequency was reduced to 8 kHz and the speech samples were preprocessed. The final step was the upsampling back to 16 kHz. Furthermore, another version of each test sample was generated using an additional level

adjustment with the P.56 software in order to compensate for the slight loudness decrement caused by the preprocessing stage. Four reference sample pairs were also provided by choosing one male and one female sentence and processing them deliberately in such a manner that they had much lower quality than the original sentences. This was done with the preprocessing function using a random mask with 30 – 60 % of the coefficients set to zero. Furthermore, the smoothing window was not in use in two of the reference samples (see Table 5.3 for the detailed descriptions). Each processed sentence in the test had the corresponding direct connection version as the quality reference. The pairs were played in random order through high-quality headphones.

Altogether 24 naive listeners participated in the test. The term naive refers to people who are not known to possess any special skills in discriminating or rating the processed speech samples. Naive listeners are very commonly used as the subjects in the listening tests because they are considered to represent the target public of the applications of speech coding. In this test, the listeners and the 24 actual test sample pairs were divided into three groups. The four reference pairs were common to all groups. Thus, each sample pair was listened by eight persons except for the reference pairs that were listened by all the listeners. In Table 5.2, the three listener groups are combined and the average grades for the processed samples with and without the additional level adjustment are shown. Table 5.3 presents the average scores given to the four references. Both tables also show the 95 % confidence intervals of the scores.

Table 5.2. Results of the CCR test. Average scores for the preprocessed speech with respect to the original ± 95 % confidence intervals.

Speaker gender	Without extra level adjustment	With extra level adjustment
Female	– 0.25 ± 0.27	0.10 ± 0.21
Male	– 0.10 ± 0.26	– 0.08 ± 0.26
Total	– 0.18 ± 0.18	0.01 ± 0.17

Table 5.3. Average scores for the quality references ± 95 % confidence intervals.

	Gender	Processing	Score
1	Female	30 % zeroed, no smoothing	– 2.50 ± 0.37
2	Female	60 % zeroed, smoothing	– 2.50 ± 0.41
3	Male	47 % zeroed, no smoothing	– 1.92 ± 0.46
4	Male	47 % zeroed, smoothing	– 1.75 ± 0.33

The results show no significant altering of the speech quality due to the preprocessing. These promising—yet indicative—results will be further discussed in Chapter 6.

5.2 Evaluation with speech codecs

Before proceeding to the consolidation of the preprocessor and a speech codec, a slight modification to the calculation of the offset was yet made. The previous modifications to the offset given by Equation (3.9) have been presented in Section 4.2. However, another

small change was found to be useful in order to mitigate the zeroing of the spectral components in the certain high-energy regions where the tonality coefficient, β, has a value between 0.2 and 1. The final form of the offset in decibels was

$$\begin{aligned} o(i) &= 100, && \beta < 0.2, \\ o(i) &= 1.2(1.1-\beta)(15+i/2), && 0.2 \le \beta < 1, \\ o(i) &= 14.5+i, && \beta = 1. \end{aligned} \tag{5.3}$$

Practically, this modification did not affect the objective quality measures calculated from the preprocessed speech but it did prevent some artefacts that had previously been audible in male speech. The percentage of the masked spectral components remained at an average of 43 %.

To test how an actual speech codec performs with the preprocessed speech, an adaptive multi-rate (AMR) algebraic CELP codec [5] was used to code Finnish speech. The codec supports eight different bit rates which were all tested with both original and preprocessed speech. The performance was assessed by calculating the segmental signal-to-noise ratio (SNR) between the input and output signals of the codec. The results are presented in Subsection 5.2.1. In addition, a listening test was arranged to get an insight into the perceptual quality of preprocessed coded speech. This perceptual evaluation was based on the absolute category rating (ACR) procedure [11] and it incorporated both the AMR CELP codec introduced above and a 2.4 kb/s MELP codec [34]. The test procedure and the results are presented in Subsection 5.2.2.

5.2.1 SNR comparison

Since CELP codecs are based on approximating the waveform of the input signal and trying to reproduce it as accurately as possible at the decoder, it is reasonable to measure the performance of the codec by comparing its output to the input on the objective basis. On the contrary, for parametric codecs such as the MELP codec, the difference between the output and input waveforms does not tell much about the quality of the codec. Denoting the input and output of the codec $x(n)$ and $y(n)$, respectively, the basic form of the signal-to-noise ratio is

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1}\left[x(n)-y(n)\right]^2}. \tag{5.4}$$

The drawback of computing SNR over the complete duration of the signal is that it will be dominated by regions that have a high energy. Since speech signals are nonstationary and contain many perceptually relevant low and high energy regions, it is better to calculate SNR in shorter segments. The results presented here were calculated with the segment length of 64 samples (8 ms). In addition, the segments where the power of the input signal is smaller than a predetermined constant were excluded from the calculation of the

segmental SNR in order to improve the accuracy. The final SNR values represent the average SNRs of the valid segments.

All of the eight modes of the CELP codec, with bit rates ranging from 4.75 kb/s to 12.2 kb/s, were used to code the same 70 minutes of Finnish speech that was used in the measurements described in Section 5.1.1. Both the original and preprocessed versions of the speech material were coded and the segmental SNRs were determined. Table 5.4 presents the average SNR values (in decibels) and the improvement percentages when using the preprocessed speech instead of the original as the coder input. Practically equal results were obtained with the older version of the preprocessor that did not contain the coefficient 1.2 in the offset.

Table 5.4. Segmental SNR between the input and output of the CELP codec.

	female			**male**		
kb/s	original	preprocessed	imp-%	original	preprocessed	imp-%
4.75	3.17	3.45	**8.71**	2.75	3.01	**9.54**
5.15	3.31	3.59	**8.56**	2.91	3.19	**9.55**
5.9	3.58	3.91	**9.25**	3.34	3.66	**9.78**
6.7	3.72	4.07	**9.64**	3.52	3.88	**10.11**
7.4	4.21	4.60	**9.21**	4.43	4.82	**8.80**
7.95	3.99	4.38	**9.72**	4.08	4.46	**9.37**
10.2	4.67	5.10	**9.23**	5.21	5.70	**9.47**
12.2	4.85	5.28	**8.76**	5.52	5.98	**8.43**

It can be clearly seen from Table 5.4 that, considering the SNR values, the codec has performed better with preprocessed than with original speech even though no optimisations have been made to the codec that would support the handling of preprocessed speech.

5.2.2 ACR listening test

To confirm the promising SNR figures, another listening test was arranged, this time at the Nokia Research Center. The test consisted in the absolute category rating (ACR) [11] in which the listeners use a five-point scale to grade the quality of the samples that have been processed with the different test conditions. The grades of the scale are illustrated in Table 5.5. The average of all scores given to a particular condition yields the corresponding mean opinion score (MOS).

Table 5.5. Grades of the ACR test.

Grade	Speech quality
5	excellent
4	good
3	fair
2	poor
1	bad

The test performed within this work contained 16 different conditions (see Table 5.6). Four of them were modulated noise reference unit (MNRU) conditions which are commonly used in listening tests to make the results comparable with other results obtained at a different time or in a different laboratory [12]. The remaining conditions tested the quality of the preprocessed speech and the performance of the CELP codec and a 2.4 kb/s MELP codec [34] with both original and preprocessed speech as the input signal.

Table 5.6. Results of the ACR test.

Condition		Female	Male
01	MNRU Q = 8 dB	1.23	1.25
02	MNRU Q = 14 dB	2.02	2.14
03	MNRU Q = 20 dB	2.98	3.04
04	MNRU Q = 26 dB	3.70	3.50
05	Direct	4.16	3.66
06	Preprocessed	4.41	4.09
07	CELP 5.15 kb/s original	3.54	3.32
08	CELP 5.15 kb/s preprocessed	3.54	3.54
09	CELP 6.70 kb/s original	3.93	3.48
10	CELP 6.70 kb/s preprocessed	4.21	3.73
11	CELP 7.95 kb/s original	4.13	3.79
12	CELP 7.95 kb/s preprocessed	4.18	3.96
13	CELP 12.2 kb/s original	4.02	3.68
14	CELP 12.2 kb/s preprocessed	4.27	3.96
15	MELP 2.4 kb/s original	2.48	2.64
16	MELP 2.4 kb/s preprocessed	2.61	2.71

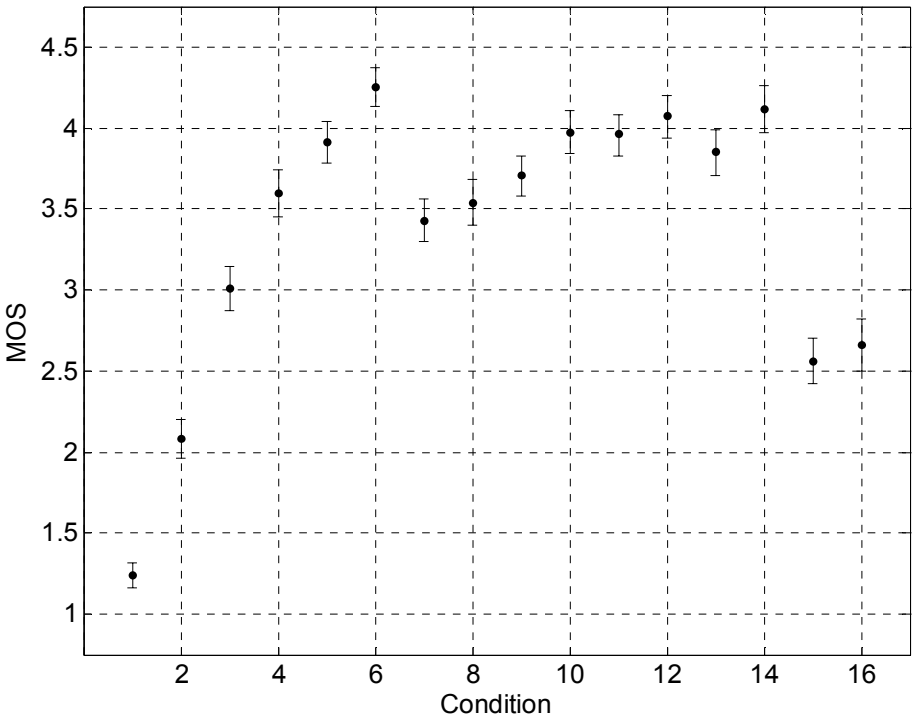

Figure 5.1. Combined MOS results with 95 % confidence intervals. The conditions are listed in Table 5.6.

The test material was spoken by two male and two female speakers and two sentences were chosen from each. These eight IRS-filtered samples were normalised to –26 dBov (i.e., relative to the overload of the digital system) [12] and processed through each of the 16 conditions. Thus, each listener assessed the quality of 128 samples in random order. Altogether 14 naive listeners participated in this test. The MOS values for male and female speakers for all conditions are shown in Table 5.6 and the combined MOS values together with their 95 % confidence intervals in Figure 5.1. The results of the ACR test are rather promising. They imply that the output speech quality of both of the tested codecs can be improved by incorporating the preprocessing technique in the front end of the codec. The results will be discussed further in the next chapter.

5.3 Summary

This chapter presented a thorough examination of the preprocessor that has been described in Chapters 3 and 4. Several objective measurements as well as two different perceptual evaluation procedures were presented. The results confirm the usefulness of the preprocessor: compared to the original speech signal, the preprocessed version has practically the same perceptual quality but it can be coded more efficiently. On the other hand, speech quality enhancement was detected when using the preprocessed speech as an input signal to parametric and waveform-approximating codecs. It was thus shown that the technique for the perceptual irrelevancy removal that has been presented in this work can be used to improve the coding efficiency of the narrowband speech codecs. The main parts of the preprocessing procedure and the achieved results are also presented in [23].

6 Discussion

The previous chapter presented promising results of several experiments with the preprocessor. This chapter briefly discusses the obtained results and considers some techniques of further improving the proposed preprocessing method. The development ideas concentrate mainly on the combination of the preprocessor and a speech codec since this kind of a system is an important future research topic and also a natural continuation of the current work.

The masking model utilised in the preprocessor often judges even more than 40 % of the frequency components of a frame to be zeroed. Nevertheless, the results of the perceptual evaluation using the comparison category rating indicate that the preprocessing causes very little or no perceptual degradation. Even without the extra speech level amplification, the quality deterioration is hardly perceivable. When combined with the additional level adjustment, the speech quality remains essentially unaltered in the preprocessing. The MOS difference between the direct connection and the preprocessed speech (conditions 5 and 6) in the ACR test even implies a slight enhancement in the speech quality due to the preprocessing. However, the magnitude of this particular difference in Table 5.6 should only warily be compared with the results of the earlier listening test because of the differences between the test methods. First, the scales differ from each other in both the meaning of the grades and in the resolution. Second, the masking threshold was adjusted before the latter perceptual evaluation, which may have had a slight contribution to the results. Finally, neither the number of listeners nor the test sentences were the same in the two tests. To enhance the reliability of the listening tests, they could be repeated with a larger number of listeners and test conditions. However, the indicative results obtained in this work already show the promising line of development.

The results of the ACR test indicate that the usage of the preprocessor in the front end of a speech coder systematically improves the speech quality. Both waveform-approximating and parametric codecs were tested and the direction of the change remained the same. It should be noted that these results were obtained without any modifications to the standardised speech codecs. Furthermore, the adjustment of the speech level was performed already before the encoding, contrary to the idea of using an adaptive control system at the decoder (discussed in Section 5.1.1). Consequently, the benefit provided by the reduction of energy due to the preprocessor was not exploited to the maximum. Even

better performance can be anticipated with the adaptive level control system, as well as with appropriate optimisations of the codecs, such as the retraining of the codebooks for the line spectral frequencies (LSFs) and other parameters. However, optimising a speech codec for the preprocessed speech can be somewhat complicated and may require several perceptual evaluation procedures before the final decisions can be made. In the scope of this thesis, the optimisation of a total codec was considered too time-consuming and is therefore examined only briefly on a theoretical level.

In principle, the retraining of the LSF codebooks is straightforward but it requires lots of computation if, for example, all the eight modes of the AMR CELP with their varying quantisation procedures are to be included. Furthermore, when adjusting the pitch estimation, it should be examined whether the pitch information can be extracted with the same accuracy from the original and from the preprocessed signals. If necessary, both versions of the signal could be inputs to the encoder to ensure optimal operation. This does not significantly increase the total complexity of the system since both the original and the preprocessed signal are readily available at the encoder. The final set of modifications naturally depends on the chosen codec as well as on the speech quality and the bit rate that are sought. Further considerations of this topic will be left to the future research.

After implementing the preprocessor in C language, it could ultimately be embedded in a speech encoder. This would obviate the need to process through the entire signal twice, first in the preprocessor and then in the codec. Instead of this time-consuming procedure, the preprocessed speech frames could be delivered to the encoding process as soon as they are finished by IFFT. The integration could also reduce the algorithmic delay of the combination in comparison to the delay caused by the cascade of the separate systems.

7 Concluding remarks

The exploitation of the psychoacoustic principles can ultimately lead a speech coding process into a perceptually optimal state in which the signal quality remains high despite a considerable reduction of the bit rate. Especially at low bit rates, it is very advantageous to avoid coding the perceptually irrelevant information. Methods for exploiting the properties of the human auditory system in speech and audio coding have been under increasingly active research during the last decade. The applications range from quantisation noise control and speech enhancement systems to objective signal quality measures. The common feature in the applications is the utilisation of masking, a phenomenon that is present in all real-world auditory signals. Masking means the process by which the perception of one sound is suppressed by another, louder sound, and it can be regarded as an unavoidable consequence of the limited frequency resolution of the human auditory system. The masking phenomenon has a key position also in the work described in this thesis.

The application proposed in this work was aimed at reducing the perceptual irrelevancy from narrowband speech signals in order to enable more efficient coding. The objective was to develop a generic processing block that removes the masked components from the speech signal before encoding. Applications where speech modified in this manner is used to replace the original speech at the input of a speech encoder had not been found from prior literature. The amount of masking caused by a short segment of speech was assessed by the masking threshold that was calculated using an auditory model. The threshold was compared with the spectrum of the input speech segment to detect the perceptually irrelevant simultaneously masked frequency components of the signal. The masked components were removed by adaptive filtering and the output, referred to as the preprocessed speech signal, was constructed by an overlap-add procedure.

The effects of the perceptual preprocessing on speech coding were studied through both objective measurements and perceptual evaluation. The compressibility of the preprocessed speech was examined through measures of energy and entropy, and the results indicated that somewhat more efficient compression can be applied to the speech signal after it has been preprocessed by the proposed method. To get an insight into the

perceptual quality of the preprocessed speech, an informal listening test was conducted in which the subjects were asked to compare the preprocessed speech to the original. According to the results of this test, no substantial deterioration of the speech quality was perceived even though the preprocessor often deemed more than 40 % of the spectral components of a frame to be removed. The evaluation of the proposed method was continued in a system where the preprocessor was used in the front end of a speech codec. The segmental SNR was computed between the output and the input of a standardised adaptive multi-rate CELP codec with each of its bit rates. Based on the SNR values, a notable improvement in the codec performance was achieved when the preprocessed speech was used instead of the original as the input signal of the codec. The segmental SNR increased by almost 10 % with both female and male speech. A perceptual evaluation also implied that the quality of the coded speech can be improved by using the preprocessor in front of the encoder. This evaluation procedure involved both the CELP codec mentioned above and a MELP codec with a bit rate of 2.4 kb/s. Tests with the MELP codec and with the different bit rates of the CELP codec showed consistent results.

The proposed method of reducing the perceptual irrelevancy of speech signals was thus found to be useful in narrowband speech coding. Due to the removal of the irrelevant components, the modified speech was applicable for more efficient coding while its perceptual quality remained essentially unaltered. In a sense, good speech quality was anticipated because the removal of frequency components was performed exactly at those parts of the spectrum that were imperceptible for the human ear. The rather violent technique of removing the masked frequencies might have caused audible distortions to the preprocessed speech, but this was successfully avoided by smoothing the frequency response of the adaptive filter. When the preprocessed speech was tested as the input signal of speech coders, improvements in the perceptual quality of the coded speech were detected with two fundamentally different speech codecs, the waveform-approximating CELP and the parametric MELP codec. These promising results were obtained even though the codecs had not been optimised for the preprocessed speech signal in any way.

Johnston's masking model was well adapted for the presented application with a small set of modifications that had experimentally been found to be advantageous either by the author or by previous researchers. All the functions were implemented in Matlab, but if faster operation is desired during future development work, the implementation can be written in, for example, C language. Even in the Matlab version of the implementation, the run time of some functions could be slightly reduced by appropriate optimising. In addition, further adjustment of the masking model itself, to make it work even better in the application presented in this work, is well worth considering. The masking model of PEAQ was also examined in this work. It contains some advanced features compared to the Johnston's model, such as the time domain masking consideration and the high resolution in the Bark domain. Therefore, it was expected to yield very good results when applied to the preprocessor. A narrowband compatible version of the PEAQ model was implemented and inserted into the preprocessor to replace the Johnston's model. Counter to the expectations, the masking model of Johnston outperformed that of the PEAQ in the preprocessing application. The reason for this was the masking threshold that was clearly too high when evaluated by the PEAQ model. Removing all the spectral components below this threshold resulted in unacceptable speech quality. The methods for correcting

the situation are likely to cause rather substantial deviation of the model from its original version, and searching for the final solution is left to future research.

On the whole, having a usable method for calculating the masking threshold for narrowband speech opens up some interesting future development possibilities. One of them might be the utilisation of the masking threshold in spectrum quantisation, for example, within the waveform interpolation coder. The perceptual relevance of the different spectral components could be assessed using the information provided by the threshold, obtaining a perceptually controlled error criterion for the quantiser. Another future development possibility is the optimisation of a speech codec for preprocessed speech. Since some pleasing preliminary results have already been achieved with the existing codecs, appropriate optimisations are likely to provide additional benefit. The exact amount of the bit rate savings that are made possible by the utilisation of the preprocessor could be evaluated within the optimisation. This can be considered one of the most salient future research objectives related to the work done so far.

References

[1] Atal, B. S & Schroeder, M. R. 1979. Predictive coding of speech signals and subjective error criteria. IEEE Transactions on acoustics, speech and signal processing. Vol 27, no. 3, pp. 247–254.

[2] Beerends, J. G. & Stemerdink, J. A. 1992. A perceptual audio quality measure based on a psychoacoustic sound representation. Journal of the Audio engineering society. Vol 40, no. 12, pp. 963–978.

[3] Cave, C. R. 2002. Perceptual modelling for low-rate audio coding. Master of Engineering thesis. Montreal, McGill University, Department of Electrical & Computer Engineering. 86 p.

[4] Edler, B. & Schuller, G. 2000. Audio coding using a psychoacoustic pre- and post-filter. Proceedings of the 2000 IEEE International conference on acoustics, speech, and signal processing, Istanbul, Turkey, June 5–9, IEEE, pp. 881–884.

[5] Ekudden, E., Hagen, R., Johansson, I. & Svedberg, J. 1999. The adaptive multi-rate speech coder. Proceedings of the 1999 IEEE Workshop on speech coding, Porvoo, Finland, June 20–23, pp. 117–119.

[6] ISO/IEC 11172-3. 1993. Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio. ISO/IEC. 150 p.

[7] ISO/IEC 13818-7. 1997. Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC). BSI ISO/IEC. 190 p.

[8] ISO/IEC 14496-3. 2001. Information technology – Coding of audio-visual objects – Part 3: Audio. ISO/IEC. 974 p.

[9] ITU-R BS.1387. 1998. Method for objective measurements of perceived audio quality. ITU. 100 p.

[10] ITU-T Recommendation P.56. 1994. Objective measurement of active speech level. ITU. 12 p.

[11] ITU-T Recommendation P.800. 1996. Methods for subjective determination of transmission quality. ITU. 28 p.

[12] ITU-T Recommendation P.810. 1996. Modulated noise reference unit (MNRU). ITU. 9 p.

[13] ITU-T Software tool library manual. 1996. ITU. Geneva. 153 p.

[14] Jayant, N., Johnston, J. D. & Safranek, R. J. 1993. Signal compression based on models of human perception. Proceedings of the IEEE. Vol 81, no. 10, pp. 1385–1422.

[15] Johnston, J. D. 1988. Transform coding of audio signals using perceptual noise criteria. IEEE Journal on selected areas in communications. Vol 6, no. 2, pp. 314–323.

[16] Kabal, P. 2002. An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality. Montreal, Department of Electrical & Computer Engineering, McGill University, TSP Lab technical report. 89 p.

[17] Kim, N. S. & Chang J–H. 2002. A preprocessor for low-bit-rate speech coding. IEEE Signal processing letters. Vol 9, no. 10, pp. 318–321.

[18] Kleijn, W. B. & Haagen, J. 1995. Waveform interpolation for coding and synthesis. Kleijn, W. B., Paliwal, K. K. (eds.). Speech coding and synthesis. Amsterdam, Elsevier Science B. V. pp. 175–207.

[19] Lukasiak, J., Burnett, I. S., Chicharo, J. F. & Thomson, M. M. 2000. Linear prediction incorporating simultaneous masking. Proceedings of the 2000 IEEE International conference on acoustics, speech and signal processing, Istanbul, Turkey, June 5–9, IEEE, pp. 1471–1474.

[20] Lukasiak, J. & Burnett, I. S. 2000. Exploiting simultaneously masked linear prediction in a WI speech coder. Proceedings of the 2000 IEEE Workshop on speech coding, Delavan, WI, USA, September 17–20, IEEE, pp. 11–13.

[21] Lukasiak, J., Burnett, I. S. & Ritz, C. H. 2001. Low rate speech coding incorporating simultaneously masked spectrally weighted linear prediction. Proceedings of the 2001 Eurospeech, Aalborg, Denmark, September 3–7, ISCA, pp. 1989–1992.

[22] Lukasiak, J. & Burnett, I. S. 2001. Source enhanced linear prediction of speech incorporating simultaneously masked spectral weighting. Journal of telecommunications and information technology. Vol 2001, no. 3, pp. 15–23.

[23] Lähdekorpi, M., Nurminen, J., Heikkinen, A. & Saarinen, J. 2003. Perceptual irrelevancy removal in narrowband speech coding. Submitted to Eurospeech 2003, Geneva, Switzerland, September 1–4, IDIAP.

[24] Moore, B. C. J. 1995. Hearing. 2nd ed. San Diego, Academic press. 468 p.

[25] Moore, B. C. J. 1997. An introduction to the psychology of hearing. 4th ed. London, Academic press. 373 p.

[26] Najafzadeh-Azghandi, H. & Kabal, P. 1999. Improving perceptual coding of narrowband audio signals at low rates. Proceedings of the 1999 IEEE International conference on acoustics, speech and signal processing, Phoenix, AZ, USA, March 15–19, IEEE, pp. 913–916.

[27] Nurminen, J. 2001. Pitch-cycle waveform quantization in a 4.0 kbps WI speech coder. Master of Science thesis. Tampere, Tampere University of Technology, Department of Information Technology. 98 p

[28] Painter, T. & Spanias, A. 1997. A review of algorithms for perceptual coding of digital audio signals. Proceedings of the 1997 13th International conference on digital signal processing, Santorini, Greece, July 2–4, IEEE, pp. 179–208.

[29] Painter, T. & Spanias, A. 2003. Sinusoidal analysis-synthesis of audio using perceptual criteria. EURASIP Journal on applied signal processing. Vol 2003, no. 1, pp. 15–20.

[30] Paliwal, K. K. & Kleijn, W. B. 1995. Quantization of LPC parameters. Kleijn, W. B., Paliwal, K. K. (eds.). Speech coding and synthesis. Amsterdam, Elsevier Science B. V. pp. 433–466.

[31] Parsons, T. W. 1987. Voice and speech processing. New York, McGraw-Hill. 402 p.

[32] Rass, U. & Steeger, G. H. 1999. Reducing time domain aliasing in adaptive overlap-add algorithms. Proceedings of the 138th Meeting of the Acoustical society of America, Columbus, OH, USA, November 1–5, ASA.

[33] Schroeder, M. R., Atal, B. S. & Hall, J. L. 1979. Optimizing digital speech coders by exploiting masking properties of the human ear. Journal of the Acoustical society of America. Vol 66, no. 6, pp. 1647–1652.

[34] Supplee, L. M., Cohn, R. P. & Collura, J. S. 1997. MELP: The new federal standard at 2400 bps. Proceedings of the 1997 IEEE International conference on acoustics, speech, and signal processing, Munich, Germany, April 21–24, IEEE, pp. 1591–1594.

[35] Terhardt, E. 1979. Calculating virtual pitch. Hearing research. Vol 1, pp. 155–182.

[36] Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K. & Feiten, B. 2000. PEAQ – The ITU standard for objective measurement of perceived audio quality. Journal of the Audio engineering society. Vol 48, no. 1/2, pp. 3–27.

[37] Thiemann, J. 2001. Acoustic noise suppression for speech signals using auditory masking effects. Master of Engineering thesis. Montreal, McGill University, Department of Electrical & Computer Engineering. 74 p.

[38] Thiemann, J. & Kabal, P. 2002. Low distortion acoustic noise suppression using a perceptual model for speech signals. Proceedings of the 2002 IEEE Workshop on speech coding, Tsukuba, Japan, October 6–9, pp. 172–174.

[39] Tsoukalas, D. E., Paraskevas, M. & Mourjopoulos, J. N. 1993. Speech enhancement using psychoacoustic criteria. Proceedings of the 1993 IEEE International conference on acoustics, speech and signal processing, Minneapolis, MN, USA, April 27–30, IEEE, pp. 359–362 (Vol 2).

[40] Tsoukalas, D. E., Mourjopoulos, J. N. & Kokkinakis G. 1997. Speech enhancement based on audible noise suppression. IEEE Transactions on speech and audio processing. Vol 5, no. 6, pp. 497–514.

[41] Verma, T., Bilbao, S. & Meng, T. H. Y. 1996. The digital prolate spheroidal window. Proceedings of the 1996 IEEE International conference on acoustics, speech, and signal processing, Atlanta, GA, USA, May 7–10, IEEE, pp. 1351–1354.

[42] Virag, N. 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Transactions on speech and audio processing. Vol 7, no. 2, pp. 126–137.

[43] Zwicker, E. & Fastl, H. 1990. Psychoacoustics. 1st ed. Berlin, Springer. 354 p.

Appendix A

Critical band numbers and the corresponding frequency limits in hertz, as presented in [43, p. 142].

Band number	Lower edge (Hz)	Centre (Hz)	Upper edge (Hz)
0	0	50	100
1	100	150	200
2	200	250	300
3	300	350	400
4	400	450	510
5	510	570	630
6	630	700	770
7	770	840	920
8	920	1000	1080
9	1080	1170	1270
10	1270	1370	1480
11	1480	1600	1720
12	1720	1850	2000
13	2000	2150	2320
14	2320	2500	2700
15	2700	2900	3150
16	3150	3400	3700
17	3700	4000	4400
18	4400	4800	5300
19	5300	5800	6400
20	6400	7000	7700
21	7700	8500	9500
22	9500	10500	12000
23	12000	13500	15500
24	15500		