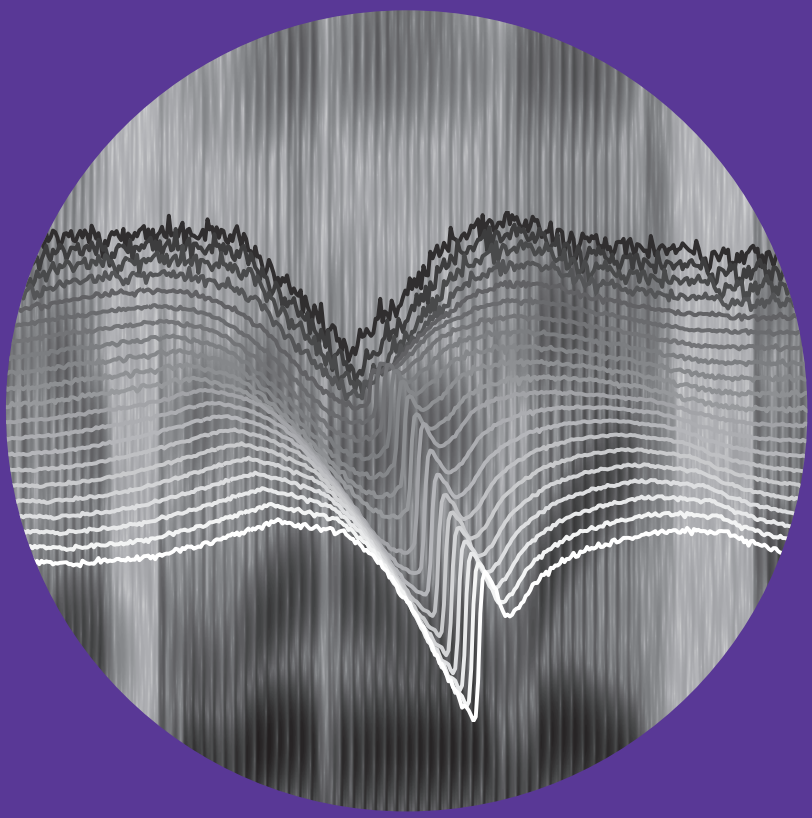
Department of Signal Processing and Acoustics

Voice source modelling techniques for statistical parametric speech synthesis

Tuomo Raitio



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 40/2015

Voice source modelling techniques for statistical parametric speech synthesis

Tuomo Raitio

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall S1 of the school on 5 June 2015 at 12.

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

Supervising professor
Professor Paavo Alku

Thesis advisor
Professor Paavo Alku

Preliminary examiners
Professor Steve Renals, University of Edinburgh, UK
Assistant Professor Jon Gudnason, Reykjavik University, Iceland

Opponent
Professor Yannis Stylianou, University of Crete, Greece

Aalto University publication series
DOCTORAL DISSERTATIONS 40/2015

ISBN 978-952-60-6136-8 (printed)
ISBN 978-952-60-6137-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6137-5

Unigrafia Oy
Helsinki 2015

Finland

441 697
Printed matter

Aalto University

Abstract

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

Author
Tuomo Raitio

Name of the doctoral dissertation
Voice source modelling techniques for statistical parametric speech synthesis

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 40/2015

Field of research Speech and language technology

Manuscript submitted 16 December 2014 **Date of the defence** 5 June 2015

Permission to publish granted (date) 5 March 2015 **Language** English

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

Abstract

Speech is the most natural way of human communication, and thus designing a machine that imitates human speech has long fascinated people. Only rather recently, due to digitisation of speech and increase in computing power, this goal has become feasible. Although speech synthesis is used today in various applications from human-computer interaction to assistive technologies, the performance of modern speech synthesisers is far from the abilities of human speakers.

The ultimate goal of text-to-speech (TTS) synthesis is to read any text and convert it to intelligible and natural sounding speech with the desired contextual and speaker characteristics. Meeting all of these goals at once makes this task extremely difficult. Moreover, the quality of the speech signal cannot be compromised since humans are very sensitive in perceiving even the slightest artefacts in a speech signal.

This thesis aims to improve both the naturalness and expressivity of speech synthesis by developing speech processing algorithms that utilise information from the speech production mechanism. One of the key algorithms in this work is glottal inverse filtering (GIF), which is used for estimating the voice source signal from recorded speech. The voice source is known to be the origin of several essential acoustic cues used in spoken communication, such as the fundamental frequency, but it is also related to acoustic cues underlying voice quality, speaking style, and speaker identity, all of which contribute to the naturalness and expressivity of speech. Accurate modelling of the voice source is often overlooked in conventional speech processing algorithms, and this work aims to improve especially this shortcoming.

In this thesis, two new GIF methods are first proposed that can be used for improved estimation of the voice source signal. Secondly, several novel voice source parameterization and modelling methods are developed that can be used in statistical parametric speech synthesis (SPSS) to improve naturalness and expressivity. Thirdly, using GIF and the voice source modelling methods in the context of SPSS, expressive voices are created that aim to cover various human speaking styles used in everyday spoken communication. Finally, the created synthetic voices are assessed using extensive subjective evaluation in different listening conditions. The results of the evaluation show that the naturalness and expressivity of synthetic speech can be enhanced using the techniques proposed in this thesis, and that the voices are perceived to be more suitable in various realistic contexts. Thus, the methods presented in this thesis provide a large potential to enhance the naturalness, expressivity, and suitability of speech synthesis in various applications.

Keywords statistical parametric speech synthesis, voice source modelling, glottal inverse filtering, voice quality, expressive speech synthesis

ISBN (printed) 978-952-60-6136-8 **ISBN (pdf)** 978-952-60-6137-5

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2015

Pages 287 **urn** http://urn.fi/URN:ISBN:978-952-60-6137-5

A! Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Tuomo Raitio

Väitöskirjan nimi
Puheen äänilähteen mallintaminen tilastollisessa parametrisessa puhesynteesissä

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 40/2015

Tutkimusala Puhe- ja kieliteknologia

Käsikirjoituksen pvm 16.12.2014 **Väitöspäivä** 05.06.2015

Julkaisuluvan myöntämispäivä 05.03.2015 **Kieli** Englanti

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

Tiivistelmä

Puhe on ihmisten luonnollisin tapa kommunikoida, ja siksi puhetta tuottavan koneen suunnittelu on jo kauan kiehtonut ihmisiä. Kuitenkin vasta viime vuosikymmeninä puhesynteesistä on tullut käytännössä mahdollista, mikä suureksi osaksi on johtunut puheen digitaalisesta esitysmuodosta ja kasvaneesta laskentatehosta. Vaikka puhesynteesiä käytetään nykyään monenlaisissa sovelluksissa, kuten ihmisen ja tietokoneen vuorovaikutuksessa sekä avustavassa teknologiassa, nykyiset puhesyntetisaattorit ovat kuitenkin vielä kaukana ihmisten monipuolisesta puheentuottokyvystä.

Puhesynteesin perimmäinen tavoite on muuttaa mikä tahansa teksti ymmärrettäväksi ja luonnollisen kuuloiseksi puheeksi, josta välittyvät myös tilanteeseen sopivat ja puhujalle ominaiset puheen piirteet. Näiden kaikkien tavoitteiden saavuttaminen yhtä aikaa on erittäin haastavaa, minkä lisäksi puhesignaalin laatu pitää olla erittäin hyvä, koska ihminen on hyvin herkkä havaitsemaan pienimpiäkin virheitä puhesignaalissa.

Tämän väitöskirjan tavoitteena on parantaa sekä puhesynteesin laatua että ilmaisuvoimaa kehittämällä puheenkäsittelymenetelmiä, jotka tarkemmin hyödyntävät informaatiota puheentuoton toimintatavasta. Yksi tämän työn tärkeimmistä menetelmistä onkin äänilähteen käänteissuodatus, minkä avulla äänitetystä puheesta voidaan määrittää äänilähdesignaali. Tämä signaali on erittäin tärkeä puheen havaitsemisen kannalta, sillä se vaikuttaa olennaisesti niihin akustisiin piirteisiin, jotka liittyvät ääntö- ja puhetapaan ja siten puheen persoonallisiin piirteisiin. Vaikka nämä piirteet vaikuttavat merkittävästi puheen luonnollisuuteen ja ilmaisuvoimaan, perinteisissä puhesynteesimenetelmissä käytetään yleensä hyvin yksinkertaistettua äänilähdesignaalin mallintamista. Tässä työssä pyritään parantamaan synteettisen puheen laatua keskittymällä erityisesti edellämainittuun ongelmaan.

Tässä väitöskirjassa esitetään ensiksi kaksi uutta äänilähteen käänteissuodatusmenetelmää, jotka mahdollistavat tarkemman äänilähdesignaalin määrittämisen puheesta. Toiseksi työssä esitetään useita uusia äänilähteen mallintamistekniikoita, joita voidaan käyttää tilastollisessa parametrisessa puhesynteesissä parantamaan puheen luonnollisuutta ja ilmaisuvoimaa. Kolmanneksi käyttämällä äänilähteen käänteissuodatus- ja mallintamistekniikoita työssä luotiin synteettisiä ääniä, jotka pyrkivät kattamaan erilaisia puhetyylejä. Lopuksi luodut äänet arvioitiin erilaisissa koeympäristöissä kuuntelukokein, joiden tulokset osoittavat että äänien luonnollisuus, ilmaisuvoima ja tilanteeseen sopivuus parani käyttämällä työssä esitettyjä menetelmiä. Näin ollen kehitetyt menetelmät tarjoavat huomattavan mahdollisuuden parantaa synteesin luonnollisuutta, ilmaisuvoimaa ja soveltuvuutta erilaisissa puhesynteesisovelluksissa.

Avainsanat tilastollinen parametrinen puhesynteesi, äänilähteen mallintaminen, äänilähteen käänteissuodatus, äänenlaatu, ekspressiivinen puhesynteesi

ISBN (painettu) 978-952-60-6136-8 **ISBN (pdf)** 978-952-60-6137-5

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2015

Sivumäärä 287 **urn** http://urn.fi/URN:ISBN:978-952-60-6137-5

Preface

This project started in 2007 when Prof. Paavo Alku recruited me to develop technology for speech synthesis. Originally, I had started my studies at the Helsinki University of Technology (currently Aalto University) aiming to major in signal processing and acoustics, since I had a great interest in sound and music. Due to my thesis topic, I ended up doing research on speech, which was not a bad choice after all. The first results of the project showed tremendous potential, and I started to get really interested in the topic. Despite the occasional unavoidable moments of doubt while doing PhD research, I decided to continue the journey deeper into speech synthesis. After all, the positive decision was easy, for which I am grateful to many people.

First of all, I would like to thank my supervisor Prof. Paavo Alku, who has supported my research at every stage. No matter what I have proposed to do research on, he has always given me encouraging and positive feedback and given the freedom to dive into topics that really interest me—this has kept my motivation towards this work really high. In addition to great guidance, he has been a truly invaluable source of information, especially in topics related to speech production and analysis. He has also provided me with a financially secure position, which has given the required amount of confidence and continuity to this work. Finally, he has been a great boss to me by never telling exactly what to do, but instead he has subtly shared ideas to trigger the researcher in me.

Speech synthesis has a long history at the Department of Signal Processing and Acoustics. During 1973–80 and 1987–90, Prof. Matti Karjalainen (in memoriam 1946–2010), the founder or the acoustics lab, and Prof. Unto K. Laine were developing speech synthesis by rule, a modern method at the time. I ended up being part of the next phase of speech synthesis research at Aalto University, which would not have been pos-

sible without the long-term research on speech synthesis by the speech research group at the University of Helsinki, and in particular that by Prof. Martti Vainio and my closest colleague Antti Suni. Combining the knowledge of these two groups has been the foundation of this successful research.

Most of the research happens when working on a computer, but many of the most relevant ideas and research directions emerge when sitting down with one's colleague for a discussion (or a beer). I am deeply grateful to Antti Suni for years of fruitful collaboration on this research topic. It has been a pleasure to develop and share ideas with Antti and draw inspiration from his deep knowledge in speech synthesis, especially the linguistic aspects. I also want to thank Prof. Martti Vainio for sharing his ideas, and especially for always being enthusiastic about every aspect of research in speech synthesis, which has helped this work a lot.

I am also grateful to the co-authors of the publications in this thesis: Harri Auvinen for performing the Markov chain Monte Carlo computations; Samuli Siltanen for providing ideas and contributing to the writing and mathematical equations; Brad H. Story for providing physically modelled synthetic speech data; Junichi Yamagishi for years of collaboration and sharing his expertise and ideas, Hannu Pulakka for helping me with dozens of things while getting used to the work of a researcher; Jani Nurminen for sharing ideas and giving feedback at the time the research was funded by Nokia; Lauri Juvela for running listening tests; Thomas Drugman for years of collaboration, good discussions, and sharing his expertise on several issues; Jouni Pohjalainen for helping me record the shouting database (which was pretty fun stuff); John Kane for enthusiastic collaboration and sharing ideas; and Christer Gobl for feedback to an article and a lovely Gaelic dinner with his charming family. I am also grateful to all my co-authors in publications not presented in this thesis.

I am really grateful to all the people in the acoustics lab who have provided a really unique and inspiring working environment. First, I want to thank all my current and former office roommates: Toni Hirvonen, Carlo Magi (in memoriam 1980–2008), Jouni Pohjalainen, and Lauri Juvela, for all the laughs, informal and professional discussions, and also the relaxed atmosphere where working is easy and efficient. Equally well I want to thank all the people with whom I have shared the lab, enjoyed lunch, and had various interesting discussions, and of course, played table football and built all kinds of things from Legos (thanks to Ville Pulkki). I am

grateful to you all: Mikkis, Olli S., Antti, Tapani (especially for the band), Hannu, Jouni, Manu, Magge, Okko, Symeon, Akis, Emma, Jussi R., Sofoklis, Sami, Seppo, Dhanu, Jukka (especially for the nice company during occupational and recreational travels), Tomppa, Reima (especially for good climbing company), Henkka, Jussi P. (especially for the mobile phone orchestra), Mairas (for Lausumo), Ville P., Unski (especially for various interesting discussions), Heikku, Marko, Henna, Javier, Juha, Miikka, Julia, Ville S., Ilkka, Teemu, Jykke, Vesa, Catarina, Olli R., Heikki, Mikko K., Julian, Alessandro, Rémi, and all the rest of the current and former lab colleagues who could not fit into this short list. I also want to thank Heidi, Hynde, Lea, Mara, Markku, Mirja, Ulla, and Tarmo for taking good care of all the practical issues in the lab. I also want to thank Prof. Simon King for collaboration and hosting me during my 6-month visit at CSTR at the University of Edinburgh. I want to thank all the colleagues and friends at CSTR, especially Tom, Rasmus, Cassia, Gustav, Oliver, and my office roommate Shinji. I also want to thank Luis Costa for proofreading my thesis, and the pre-examiners of the thesis, Prof. Steve Renals from the University of Edinburgh and Assistant Prof. Jon Gudnason from the Reykjavik University, for their valuable comments on improving the thesis. I am also grateful to the whole speech research community for the encouraging and enthusiastic research atmosphere and for organising wonderful conferences around the world.

My work has been funded by several organisation, and I am really grateful to them all for both the money and for seeing that speech synthesis is really a topic worth investing all the time and money. During the years my work has been funded by the Academy of Finland, Nokia, Graduate School at Aalto University School of Electrical Engineering, Aalto University Multidisciplinary Institute of Digitalisation and Energy, European Community's Seventh Framework Programme Simple4All, and Finnish Funding Agency for Technology and Innovation (Tekes). I have also been supported in the form of personal grants by the Nokia Foundation, Emil Aaltonen Foundation, Finnish Science Foundation for Economics and Technology (KAUTE), HPY Research Foundation, and Research and Training Foundation of TeliaSonera Finland Oyj.

They say time is money, but it is more. I have used many long days and evenings doing this research not mainly for the money, but for the interest in this fascinating topic. At the same time I have also been privileged to spend time with my wonderful friends and various inspiring persons. I am

very grateful for all the cheerful times with you. Finally, I would like to express my warmest gratitude to my parents, Eija and Matti, my brother Arimatias, and my sisters, Tuovi and Jenni, for all the joy and support over the years. Regarding the topic of this thesis, I want to express special thanks to my father and grandfather Pentti (in memoriam 1930–2014) who evoked the interest in music and sound in me, which I have been following since then in my own way.

Espoo, March 17, 2015,

Tuomo Raitio

Contents

List of publications

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

I Harri Auvinen, Tuomo Raitio, Samuli Siltanen, Brad H. Story, and Paavo Alku. Automatic glottal inverse filtering with the Markov chain Monte Carlo method. *Computer Speech and Language*, vol. 28, no. 5, pp. 1139–1155, September 2014.

II Manu Airaksinen, Tuomo Raitio, Brad Story, and Paavo Alku. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 596–607, March 2014.

III Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, January 2011.

IV Tuomo Raitio, Antti Suni, Hannu Pulakka, Martti Vainio, and Paavo Alku. Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, pp. 4564–4567, May 2011.

V Tuomo Raitio, Antti Suni, Lauri Juvela, Martti Vainio, and Paavo Alku. Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort. In *Proceedings of the*

15th Annual Conference of the International Speech Communication Association (Interspeech), Singapore, pp. 1969–1973, September 2014.

VI Thomas Drugman and Tuomo Raitio. Excitation modeling for HMM-based speech synthesis: Breaking down the impact of periodic and aperiodic components. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, pp. 260–264, May 2014.

VII Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku. Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise. *Computer Speech and Language*, vol. 28, no. 2, pp. 648–664, March 2014.

VIII Tuomo Raitio, Antti Suni, Jouni Pohjalainen, Manu Airaksinen, Martti Vainio, and Paavo Alku. Analysis and synthesis of shouted speech. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, pp. 1544–1548, August 2013.

IX Tuomo Raitio, John Kane, Thomas Drugman, and Christer Gobl. HMM-based synthesis of creaky voice. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, pp. 2316–2320, August 2013.

Author's contribution

Publication I: "Automatic glottal inverse filtering with the Markov chain Monte Carlo method"

The author developed and implemented the signal processing parts of the method, and developed the complete method together with the first author. The author also designed and implemented the objective evaluations and primarily wrote the paper, except for the descriptions on the Markov chain Monte Carlo algorithm.

Publication II: "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction"

The author was involved in the original research of the algorithm, and for this article the author implemented the proposed method as part of a speech synthesiser that was used for generating speech samples for the subjective listening tests. The author also recorded the speech/shout database used in the listening tests, and designed and wrote the parts concerning the listening tests with the first author. The author also participated in writing and editing the final article.

Publication III: "HMM-based speech synthesis utilizing glottal inverse filtering"

The author developed and implemented the proposed method, ran all the subjective listening tests, analysed the results, primarily wrote the article, and generated all figures except Figure 5.

Publication IV: "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis"

The concept of the method was developed by the author with the first co-author. The author developed and implemented the algorithm, ran the subjective evaluations, and primarily wrote the entire paper.

Publication V: "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort"

The author outlined the study, developed the method and implemented the algorithms. The author generated all the samples for the experiments with the second author and analysed the results of the subjective listening tests. The author wrote the entire article with helpful comments and corrections from the co-authors.

Publication VI: "Excitation modeling for HMM-based speech synthesis: Breaking down the impact of periodic and aperiodic components"

The author developed and designed the algorithms and tests together with the co-author, trained and built all the synthetic voices, and analysed the listening test results. The author wrote the paper together with the co-author.

Publication VII: "Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise"

The author implemented the synthesis methods for the study with the exception of the HMM training and synthesis, which was done by the second author. The author designed the subjective evaluations with the co-authors, ran all evaluations, and conducted the analysis of results. The author primarily wrote the entire paper and generated all figures except Figures 1 and 2.

Publication VIII: "Analysis and synthesis of shouted speech"

The author outlined the study, recorded the shouted and modal speech samples with the third author, and ran all the objective analyses. The author designed and ran the listening tests with the help of the fourth author. The author primarily wrote the entire paper but received helpful corrections and suggestions from co-authors.

Publication IX: "HMM-based synthesis of creaky voice"

The tests in the study were designed and conducted in collaboration with the two first co-authors, and the author conducted the experiment on fundamental frequency analysis and built all the voices used in the study. The paper was written together with the two first co-authors.

List of abbreviations

List of abbreviations appearing in the dissertation

ACR	absolute category rating
AI	articulation index
AME	attenuated main excitation
AR	autoregressive
ARMA	autoregressive moving average
ARX	autoregressive with exogenous input
ASR	automatic speech recognition
AVSS	average voice-based speech synthesis
CALL	computer-assisted language learning
CCD	complex-cepstrum-based decomposition
CCR	comparison category rating
CD-HMM	continuous-density hidden Markov model
CMOS	comparison mean opinion score
CP	closed phase
ClQ	closing quotient
CQ	closed quotient
DAP	discrete all-pole modelling
DBN	deep belief network
DCT	discrete cosine transform
DNN	deep neural network
DP	dynamic programming
DRT	diagnostic rhyme test
DSM	deterministic plus stochastic model
DSP	digital signal processing
DYPSA	dynamic programming projected phase-slope algorithm

EGG	electroglottography
EM	expectation-maximization
EMA	electromagnetic articulography
ERB	equivalent rectangular bandwidth
FIR	finite impulse response
GCI	glottal closure instant
GHNR	glottal harmonic-to-noise ratio
GIF	glottal inverse filtering
GOI	glottal opening instant
GP	glimpse proportion
GV	global variance
HCI	human-computer interaction
HMM	hidden Markov model
HNM	harmonic plus noise model
HNR	harmonic-to-noise ratio
HRF	harmonic richness factor
HSMM	hidden semi-Markov model
HTK	hidden Markov model toolkit
HTS	HMM-based speech synthesis system
IAIF	iterative adaptive inverse filtering
IFT	inverse Fourier transform
IIR	infinite impulse response
IPA	International Phonetic Association
JND	just-noticeable difference
LER	letter error rate
LF	Liljencrants–Fant (glottal flow derivative model)
LP	linear prediction
LPC	linear predictive coding
LSD	log-spectral distortion
LSF	line spectral frequency
LSP	line spectral pair
LSTM	long short-term memory
LTI	linear time-invariant
LTS	letter-to-sound
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
MDL	minimum description length
MDN	mixture density network

MGE	minimum generation error
MLLR	maximum likelihood linear regression
MLPG	maximum likelihood parameter generation
MLSA	mel-log spectrum approximation
MMI	maximum mutual information
MOS	mean opinion score
MRI	magnetic resonance imaging
MRT	modified rhyme test
MSD	multi-space probability distribution
NAQ	normalized amplitude quotient
NLP	natural language processing
OQ	open quotient
P2V	peak-to-valley
PCA	principal component analysis
PESQ	perceptual evaluation of speech quality
POS	part of speech
PSOLA	pitch-synchronous overlap-add
PSP	parabolic spectral parameter
QCP	quasi-closed phase
QOQ	quasi-open quotient
RBM	restricted Boltzmann machine
RK	Rosenberg–Klatt (glottal flow model)
RMSE	root mean squared error
RNN	recurrent neural network
RQ	return quotient
SAT	speaker-adaptive training
SEDREAMS	speech event detection using the residual excitation and a mean-based signal
SGD	speech-generating device
SII	speech intelligibility index
SLM	sinusoidal likeness measure
SNR	signal-to-noise ratio
SPL	sound pressure level
SPSS	statistical parametric speech synthesis
SPTK	speech signal processing toolkit
SQ	speed quotient
SRT	speech reception threshold
STE	short-term energy

STI	speech transmission index
STOI	short-time objective intelligibility
STRAIGHT	speech transformation and representation using adaptive interpolation of weighted spectrum
SUS	semantically unpredictable sentence
ToBI	tones and break indices
TTS	text-to-speech
VOT	voice onset time
WER	word error rate
WLP	weighted linear prediction
YAGA	yet another GCI/GOI algorithm
ZFF	zero-frequency filtering
ZZT	zeros of the z transform

List of abbreviations appearing in the attached publications

ANOVA	analysis of variance
CC	complex cepstrum
CD	continuous probability distribution
CSMAPLR	constrained structural maximum a posteriori linear regression
DQ	duration quotient
DRAM	delayed rejection + adaptive Metropolis
GLOAT	glottal analysis toolbox
HSD	honestly significant difference
ME	mixed excitation
MGC	mel-generalised cepstrum
MGLSA	mel-generalised log-spectral approximation
MSD-HSMM	multi-space probability distribution hidden semi-Markov model
MSE	mean squared error
PDF	probability distribution function
PQ	position quotient
RAPT	robust algorithm for pitch tracking
RMS	root mean square
SWIPE	sawtooth waveform inspired pitch estimator
TEMPO	time-domain excitation extractor using minimum perturbation operator

List of symbols

List of Latin symbols

$\mathbf{A}$	state transition probability distribution of an HMM
a_{ij}	state transition probability of taking a transition from state q_i to q_j
$\mathbf{B}$	output probability distribution of an HMM
$b_i(\mathbf{o})$	probability of emitting an observation $\mathbf{o}$ in state i
$c(m)$	cepstrum with quefrency index m
d_{peak}	negative peak amplitude of the glottal flow derivative
E_{e}	LF model parameter: value of the flow derivative at time t_{e}
f_0	fundamental frequency
f_{ac}	maximum amplitude of the glottal flow cycle
F_{m}	maximum voiced frequency
f_{s}	sampling frequency
$G(z)$	z transform of the glottal excitation
$L(z)$	transfer function of the lip radiation effect
$\mathbf{O}$	observation sequence or the set of alphabet of a discrete HMM
$\mathbf{o}_t$	observation vector at time t
p	all-pole model order
$\mathbf{Q}$	state sequence or the set of states of an HMM
q_t	HMM state at time t
R_{d}	LF model voice source shape parameter
$s(n)$	speech signal with time index n
$S(z)$	z transform of a speech segment
T_0	fundamental period of the glottal cycle
t_{a}	LF model parameter: return phase projection
t_{c}	closed time of the glottal cycle

t_{cl}	closing time of the glottal cycle
t_{e}	LF model parameter: time of the maximum negative value of the flow derivative
t_{o}	opening time of the glottal cycle
t_{p}	LF model parameter: time of the flow derivative zero point after maximum value
t_{ret}	return time of the glottal cycle
$V(z)$	transfer function of the vocal tract
w	linguistic specification extracted from input text
W	linguistic specification extracted from text corpus

List of Greek symbols

Δ	time derivative of a vector
λ	notation for the set of model parameters of an HMM
$\boldsymbol{\mu}$	mean vector of speech parameters
$\boldsymbol{\pi}$	initial state probability distribution of an HMM
$\boldsymbol{\Sigma}$	covariance matrix of speech parameters
ω_{g}	frequency (in radians) of the sinusoidal open phase component in the LF glottal flow derivative model

List of figures

List of tables

1. Introduction

"The limits of my language means the limits of my world"

— Ludwig Wittgenstein

"What did you do yesterday?"—Such a simple query, but how would you give an answer to this question without speech and language? Communication between people using speech and language is certainly one of the most important abilities we have, and it comes so naturally to us that we do not pay much attention to it. The richness of human language enables us to communicate arbitrarily complex meanings to other people, which to a large extent constitutes our everyday lives. We use our language for managing mundane social tasks as well as for creating and enjoying art, cherishing culture, enhancing our environment, and in developing technology and science—even our own experiences are partly governed by the inner speech.

Speech is the most natural mode of human communication, which has been the driving force underlying several significant advances in speech technology. Today, due to digital speech transmission and mobile phones, communication using speech is no longer dependent on the location of the speaker or listener. Digitisation of speech and advances in computing power and methodology have also enabled various speech technologies that imitate the human ability to speak and understand speech. Automatic speech generation (i.e., speech synthesis) and automatic speech recognition (ASR) have a wide range of existing and potential applications, and currently various international companies are heavily investing in these technologies, aiming at better services and products for their customers. At the same time, fundamental methods for these speech technologies are being developed in academia. New speech technologies are also used in various assistive technology applications to help people with disabilities. These emerging technologies have already provided an im-

pressive demonstration of future ubiquitous speech technology, which find applications wherever speech is used—in practice almost everywhere.

Clearly the performance of speech synthesis and ASR is not comparable to the human capability of producing and recognising speech. Speech synthesis and recognition are challenging problems, and there is a lot of room to improve the existing technologies. Although ASR already performs very impressively even in slightly noisy environments, speech synthesis is still lacking, for example, in the human capability of expression. The problem of speech synthesis is the vast variability of the desired output—while in ASR, the goal is to reduce variability by estimating the most probable word sequence corresponding to a spoken utterance, in speech synthesis, there are virtually infinite ways of expressing one particular word sequence. Modelling and generating speech (by computers) that has the same desired variability as human speech, in terms of, say, speaker characteristics, speaking styles, and emotions, collectively called expressivity, is a problem that still requires several technological inventions. On top of that, while modelling the expressive characteristics of human speech, the quality of the speech signal cannot be compromised since humans are very sensitive in perceiving even the slightest artefacts in speech.

This thesis aims at improving both the naturalness and expressivity of speech synthesis. This is achieved by developing speech processing algorithms that utilise information from the speech production mechanism in order to better model various perceptually important acoustic cues in speech. The voice source is known to be the origin for several essential acoustic cues used in spoken communication, such as fundamental frequency, but it is also related to acoustic cues underlying voice quality, speaking style, and speaker identity, which all contribute to the naturalness and expressivity of speech. The accurate modelling of the voice source is often overlooked in conventional speech processing algorithms, and this thesis aims at improving especially this aspect of speech synthesis.

One of the key algorithms in this work is glottal inverse filtering (GIF), which is used to estimate the voice source signal from a recorded speech signal. The voice source signal, also called the glottal flow, depicts the glottal air flow as a function of time and is further filtered by the resonances of the vocal tract cavities. By revealing this perceptually important signal, it is possible to quantify and parameterize perceptually relevant cues in the voice source and reproduce them in speech synthesis

based on the desired context. However, the estimation, parameterization, and modelling of this important signal are challenging tasks. First, considering that GIF is a difficult inverse problem, the GIF algorithm must yield accurate and robust estimates of the voice source in order to provide a useful starting point for the parameterization and modelling stages. Second, the parameterization of the glottal flow must simultaneously preserve the perceptually most relevant cues while enabling the use of statistical modelling methods in order to successfully reconstruct the excitation signal in synthesis. The synthetic voices created using glottal flow modelling must be finally evaluated by human subjects to show possible improvements or cases which require more attention and further work. By repeating this process of developing new effective speech processing algorithms combined with suitable statistical modelling and evaluation (without forgetting the development of front-end and prosody modelling), naturalness and expressivity of synthesis can be enhanced significantly, as is shown in this work.

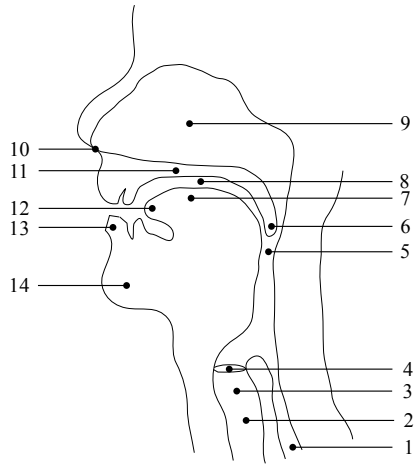
This thesis consists of two parts. In the first part of the thesis, a general overview is given on the topics relevant to this thesis, namely speech production and perception, voice source estimation and parameterization, and speech synthesis. In the second part of the thesis, the most significant publications resulting from this work are attached and sorted according to three topics. The first two publications (I, II) describe new GIF methods that improve the accuracy of voice source estimation. The following four publications (III, IV, V, VI) describe the integration of new speech processing algorithms, such as GIF and new methods for voice source modelling, into a vocoder, which is used for speech synthesis. The last three publications (VII, VIII, IX) as well as publication V use the developed speech processing algorithms and the vocoder for synthesising expressive speech, such as breathy and Lombard speech, shouted speech, and speech with a tendency for a creaky voice.

2. Speech production and perception

This section gives an overview on human speech production and perception. First, the speech production mechanism is described, concentrating on the two main vocal mechanisms and their functions in speech production: the glottal excitation and vocal tract. Also, a rough classification of the speech sounds used in spoken language is given. The function of the voice source in speech production is further elaborated by describing different voice qualities and phonation types and their functions in speech communication. A simplified speech production model, the source-filter theory, which is used in various speech processing applications, is also described. The perception of speech is also discussed by first shortly describing the hearing mechanism and its main properties, after which the acoustic cues for speech perception are highlighted. Finally, the properties of speech other than those having phonemic function are discussed, such as coarticulation, prosody, and the effect of context and the speaker.

2.1 Speech production mechanism

Speech production can be described as a result of three main components: the respiratory system, larynx, and the vocal tract. Speech is produced by exerting air from the lungs through the trachea and regulating the air flow at the larynx and the vocal tract. At the larynx, the air flow is modulated by the vocal folds, which creates the main excitation for voiced speech. The vocal tract, consisting of the pharynx, oral cavity, and nasal cavity, shapes the spectrum of the modulated air flow by creating resonances and antiresonances. The dimensions of the vocal tract can be voluntarily modified by the speaker to create various speech sounds, in which case the vocal tract acts as a time-varying filter. Speech is finally radiated through the lips and nostrils to the surrounding air as an acoustic speech



1. Oesophagus
2. Trachea
3. Larynx
4. Vocal folds
5. Pharynx
6. Uvula
7. Tongue (dorsum)
8. Oral cavity
9. Nasal cavity
10. Nostrils
11. Palate
12. Tongue tip
13. Lips
14. Jaw

Figure 2.1. Speech production mechanism.

wave. The speech production mechanism is illustrated in Figure 2.1.

The produced speech sounds can be roughly classified into two categories: voiced and unvoiced. The source of voiced speech sounds is the vibratory motion of the vocal folds, which generates a quasi-periodic excitation signal rich in harmonics. Voiced speech sounds form the main part of most West European languages (Catford, 1977). For example, 78% of phonemes in English are reported to be voiced (Catford, 1977). Unvoiced sounds are created by constricting the airflow somewhere in the vocal tract, which creates a turbulent noise source without a harmonic structure. Unvoiced sounds can be further classified according to the place and type of constriction in the vocal tract, which is discussed in more detail in Section 2.2. Many speech sounds, however, consist of both voiced and unvoiced components.

2.1.1 Glottal excitation

The respiratory system functions both as an air reservoir and as a means to exert air in a controlled manner to the upper parts of the vocal organs for the production of speech sounds. In normal inhalation and exhalation, practically no sound is emitted. In the case of voiced speech, the vocal folds are adducted using the musculature in the larynx, which results in a self-oscillating vibratory movement of the vocal folds because of the air flow. This movement results in a modulation of the air flow into small pulses. The vocal folds (or vocal cords) are two elastic ligaments in the larynx extending from the thyroid cartilage to the arytenoid cartilages, of which the latter controls the V-shaped opening between the vocal folds.

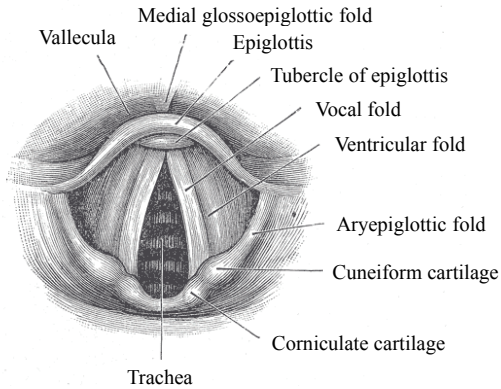


Figure 2.2. Illustration of the larynx, depicted as seen looking down from the pharynx towards trachea. The vocal folds are the two ligaments partly covering the trachea (adapted from Gray, 1918).

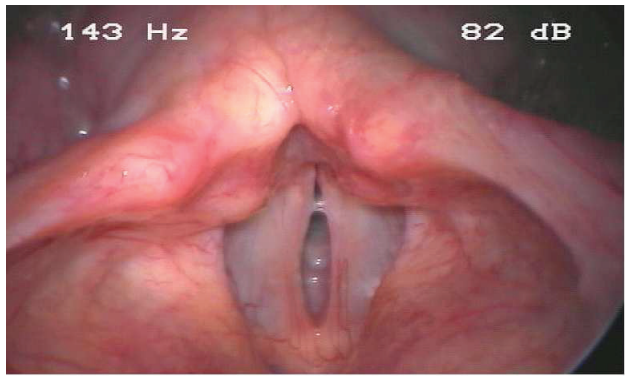


Figure 2.3. Photograph of larynx in phonation.

This opening is called the *glottis* (Flanagan, 1972a), which is illustrated in Figures 2.2 and 2.3. The behaviour of the vocal folds in phonation is illustrated in Figure 2.4. Unlike voluntary muscle movement, vocal fold vibration is a result of both aerodynamics and the elasticity of the vocal folds (van den Berg et al., 1957). First, the subglottal pressure is increased as air from the lungs is pushed upwards to the closed glottis. As the pressure builds high enough, it forces the vocal folds to open gradually. The air flow is increased, which creates an underpressure between the vocal folds, which in turn with the elastic force of the vocal folds draws the vocal folds back together. The main excitation of voiced speech is generated when the vocal folds hit together, which is seen as a strong negative peak in the glottal flow derivative signal. After the glottal closure, the subglottal pressure begins to increase again, starting a new glottal period. Since the upper and lower parts of the vocal folds have different elasticity, there is a small time lag between their movements during the

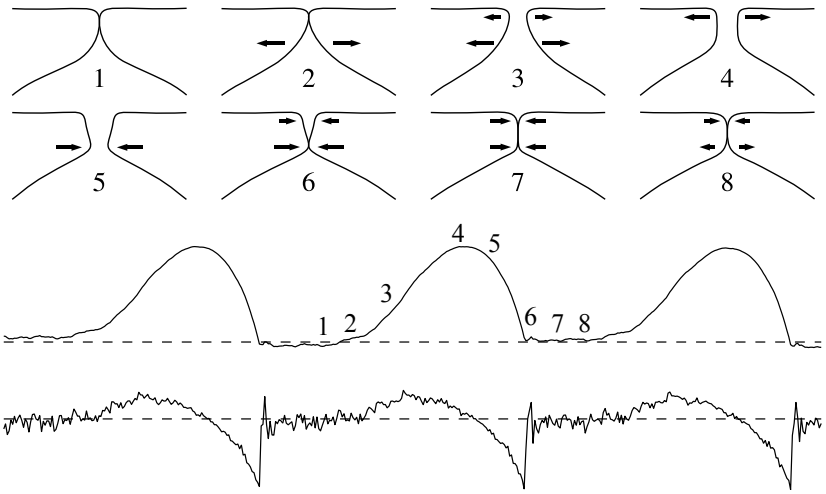


Figure 2.4. Behaviour of the vocal folds in phonation. The uppermost figure depicts the different phases of the glottal flow vibration, shown as a horizontal-frontal cross-section of the vocal folds. The two signals below represent the corresponding glottal volume-velocity waveform (upper) and its derivative (lower) (uppermost figure adapted from Story, 2002; Pulakka, 2005).

glottal cycle (O'Shaughnessy, 2000). The periodic air flow generated by the vibrating vocal folds is called the glottal flow, glottal volume-velocity waveform, or simply the voice source. The glottal flow is approximately proportional to the glottal open area, and the glottal flow signal resembles a half sine wave with a smooth opening and more abrupt closing.

The tension of the vocal folds can be controlled voluntarily by the speaker to control the pitch and phonation type of speech. The rate at which the vocal folds vibrate defines the fundamental frequency (f_0) or the pitch of speech. The pitch is determined by the mass, length, and tension of the vocal folds relative to the glottal air volume velocity. Increased vocal fold length and tension increase f_0 while increased mass decreases f_0 (Mathieson, 2000). For example, Peterson and Barney (1952) report that the average f_0 is around 132 Hz for males, 223 Hz for females, and 264 Hz for children. In an arbitrary utterance, f_0 can vary from as low as 30 Hz up to more than 600 Hz, and a soprano singer can reach an f_0 of over 1300 Hz (Titze, 1994, p. 176).

The position and tension of the vocal folds, controlled by the laryngeal muscles, and amount of airflow also determine the type of phonation. A stronger vocal fold tension will result in a more abrupt glottal closure and thus a louder and more tense speech will be produced. Different phonation types are discussed in more detail in Section 2.3.

In addition to the periodic glottal excitation, there is always some degree of aperiodicity present in the voice source signal. The aperiodicity in the voice source may stem from various phenomena, such as jitter, shimmer or waveshape change, or additive aspiration noise originating from the glottis or above it (Rothenberg, 1974; Murphy, 1999; O'Shaughnessy, 2000). Especially with female speech, the higher harmonics of speech are covered by strong aspiration noise due to the incomplete closure of the glottis (Hanson, 1995). These small deviations in the quasi-periodic glottal excitation may be a likely reason for the natural character of speech that, for example, speech synthesisers have difficulties replicating. Altogether, the various properties of the glottal flow (in addition to cues from the vocal tract) leave a specific signature on a voice, giving cues for recognising gender, age, and the speaker (Klatt and Klatt, 1990; Childers and Lee, 1991) and possible special traits or pathologies in a voice (Gómez-Vilda et al., 2009), which are all important properties when creating personalised expressive synthetic voices.

2.1.2 Vocal tract

The vocal tract consists of the pharynx and the oral and nasal cavities, which together shape the spectrum of the glottal excitation and create different speech sounds. The shape of the oral cavity can be adjusted by moving, the larynx, tongue, cheek, and lips, which results in different resonant effects called *formants*. The formant structure is used to distinguish one phoneme from another. The length of the pharynx can be also slightly modified by raising or lowering the larynx. Also, by raising or lowering the soft palate, the air flow through the nasal cavity can be controlled. The vocal tract can be considered as a single tube extending from the vocal folds to the lips, with a side branch to the nasal cavity. The cross-sectional area of the vocal tract mostly defines the resonant effects, but different speech sounds can be also created by constructing the air flow at some point in the vocal tract. An illustration of the profile of the vocal tract and the resulting vocal tract spectral envelope in phonation of three different vowels is show in Figure 2.5.

The dimensions and shape of the vocal tract vary across males, females, and children, and they also vary from person to person (Fant, 1960; Peterson and Barney, 1952; Hillenbrand et al., 1995; Story et al., 1995). Therefore, the resonant structure, and thus the resulting spectral characteristics of speech, are slightly different for each person. The personal differ-

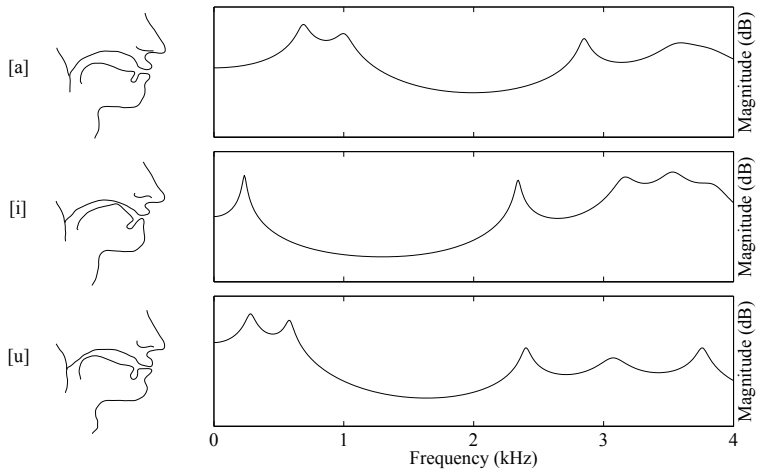


Figure 2.5. Illustration of the vocal tract shapes and the corresponding vocal tract spectral envelopes in phonation of the vowels [a], [i], and [u].

ences are smaller than the differences between different speech sounds, but they (in addition to cues from the voice source) give each speaker a characteristic timbre from which humans recognise speakers from each other and can yield lots of information about, for instance, the gender or age of a speaker.

2.2 Classification of speech sounds

Speech sounds can be classified in various ways. The classification of speech into voiced and unvoiced categories based on the speech production mechanism was already briefly described in Section 2.1. In this section, the classification of speech sounds is elaborated further with a linguistic perspective.

The simplest classification divides speech sounds into two groups: vowels and consonants. Vowels are voiced sounds that are produced by unrestricted airflow in the vocal tract. Different vowels are determined by their formant structure, primarily by their first and second formants. Vowels can be assumed to be quasi-stationary for short periods of time (e.g., 25 milliseconds) since the movement of the articulations is relatively slow. Vowels usually have a rather high energy that is concentrated at the low frequencies. The number of vowels in spoken language slightly differs depending on language and definition. The classification of vowels based on the open-closed and front-back dimensions of the vocal tract by the

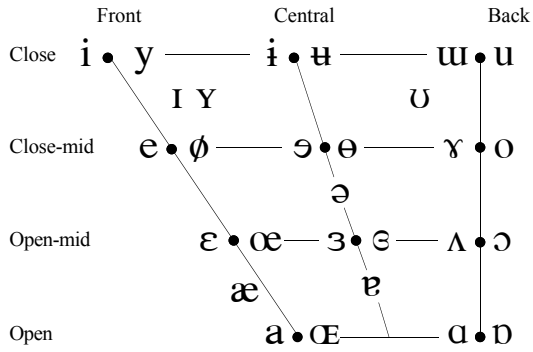


Figure 2.6. IPA classification of vowels. Where symbols appear in pairs, the one to the right represents a rounded vowel (from International Phonetic Association, 2005).

International Phonetic Association (IPA) is shown in Figure 2.6. The classification information can be utilised, for example, in speech synthesis for clustering similar speech sounds for handling data sparsity.

Consonants can be either voiced or unvoiced, but they consist of a complete or partial closure somewhere in the vocal tract. Many of the consonants involve a sudden change in the articulators, due to which the quasi-stationarity assumption as in the case of vowels, does not always hold. Many consonants are also characterised by a lower energy that is concentrated at the high frequencies. However, voiced consonants have many characteristics similar to vowels. Consonants can be further classified according to the place and manner of articulation or voicing. For example, the following classification can be used to categorise consonants:

- Plosives (e.g., [k], [p], [t], [g], [b], [d]) are produced by completely blocking the air flow at some point in the vocal tract and then suddenly releasing the air flow. Plosives are characterised by a pause, which is followed by a transient noise burst.
- Fricatives (e.g., [f], [h], [s]) are produced by forming a constriction at some point in the vocal tract so that the air flow becomes turbulent, which produces a relatively long noise-based excitation signal.
- Nasals (e.g., [m], [n]) are produced by opening the nasal tract by lowering the velum and closing the oral tract. Nasal sounds are characterised by antiformants in addition to formants. Antiformants (i.e., spectral valleys) are created since the closed oral cavity traps speech energy at certain frequencies.
- Approximants (e.g., [ʊ] [l], [ɹ], [j]) usually involve a partial constriction

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Figure 2.7. IPA classification of pulmonic consonants. Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible (from International Phonetic Association, 2005).

of the vocal tract but does not create a significant amount of turbulent noise. Due to the partial constriction, antiformants are also produced (Kent and Read, 1992).

- Other consonants, such as taps and flaps, where an articulator is rapidly hit against another, or trills (e.g., [r]), where the air flow causes an articulator to vibrate.

A detailed classification of pulmonic (produced by air pressure from the lungs) consonants by IPA is shown in Figure 2.7. In addition to pulmonic consonants, there are also non-pulmonic consonants, such as ejectives, implosives, and click consonants, but they are rather rarely used in languages. Similar to the classification of vowels, the classification of consonants can be utilised, for example, in speech synthesis for clustering similar speech sounds.

2.3 Voice quality and phonation types

Humans can use their voice in multiple different modes depending on the configuration of the glottis and the amount air pressure generated by the respiratory system. There are multiple ways to characterise and classify different voice qualities (for example, see Abercrombie, 1967; Ladefoged, 1971; Laver, 1980; Gobl, 1989; Klatt and Klatt, 1990; Gobl and Ní Chasaide, 1992; Gordon and Ladefoged, 2001). Based on a simplified functioning of the larynx, voice quality can be described as a continuum that is determined by the aperture between the arytenoid cartilages (Ladefoged, 1971). The continuum spans from whisper, whispery, and breathy speech to modal and finally to tense and creaky voice, as the aperture between the vocal folds is decreased. In the following, a more detailed overview

of different speech qualities is given in the order of the aforementioned continuum. However, modal speech is described first since it is the most common phonation mode to which all other voice qualities are compared.

Virtually all speech employs the *modal* register[1], for which the functioning of the vocal folds was already illustrated in Section 2.1.1. In modal speech, the vocal folds function in an efficient manner using moderate adduction and achieving complete glottal closure. As a result, modal speech is rather quasi-periodic with only small amounts of perturbations or aspiration noise.

In *whisper* and *whispery* voice, the vocal folds form a small triangular opening between the arytenoid cartilages, which results in a strong aspiration component. In whisper, the vocal folds do not vibrate, and thus the only excitation is the turbulent noise component. Whispery voice, however, contains also glottal vibrations, but the triangular opening makes the glottal closure incomplete, and therefore it contains both a relatively weak voicing component and a strong aspiration noise component.

In *breathy* phonation, the vocal folds vibrate in a less efficient manner than in modal phonation, and due to the greater air flow in comparison to the glottal flow vibration, there is also a strong aspiration component. The glottal closure is often incomplete in breathy voice, which results in a reduced excitation peak at the instant of glottal closure. Therefore, breathy voice is characterised by a steep spectral tilt, emphasising the low frequency components.

If the adduction is higher than in modal speech, a *tense* voice quality is produced. Due to the higher tension in the vocal folds, the air pulses through the glottis become shorter and the glottal closures become more abrupt. Therefore, tense voice is characterised by a decreased spectral tilt, but the phonation is still rather quasi-periodic. Breathy, modal, and tense speech, and their glottal and spectral characteristics are illustrated in Figure 2.8.

Harsh voice has even higher tension than tense voice (Gobl and Ní Chasaide, 2003), and therefore the vocal folds do not vibrate in a normal periodic manner. Harsh voice has additional aperiodicity and is characterised by the perception of an unpleasant, rasping sound, caused by the irregularities.

[1]Modal speech is often also referred to as *normal* speech, but in order to avoid confusion with medically normal speech, the term modal speech is used in this work.

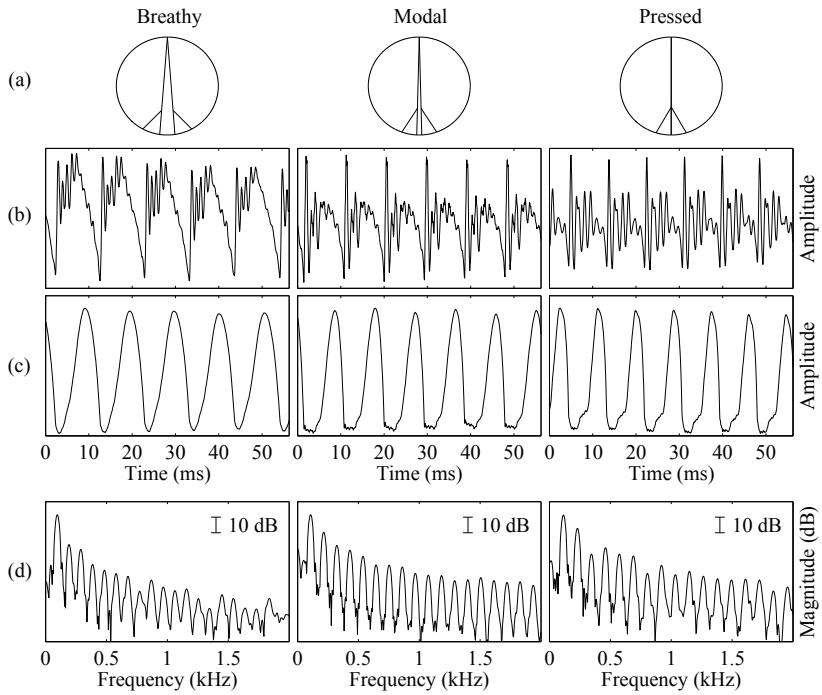


Figure 2.8. (a) Schematic illustration of the glottal configuration in breathy (left), modal (middle), and pressed (right) phonation. (b) Waveform of a sustained vowel [a] produced by a male speaker using the corresponding voice qualities. (c) Corresponding glottal flow signals estimated with glottal inverse filtering. In breathy phonation, the pulses are longer and there is no clear closed phase, whereas in modal and pressed speech, the glottal flow pulses are shorter compared to breathy phonation, and there is a clear closed phase between pulses. (d) Spectra of the estimated glottal flow signals (only shown for 0–2 kHz). The spectrum of breathy phonation shows a clear emphasis on the fundamental frequency component. The spectral envelope is also steeper, and there is more noise at the higher frequencies. In pressed phonation, the first few harmonics are emphasised in comparison to modal and breathy speech.

Creaky voice, *vocal fry*, or *laryngealisation* is a register where the vocal folds are tightly adducted and thus only a small amount of air is passed in each glottal cycle (Blomgren et al., 1998). The open phase of the glottal flow is extremely short and there is an abrupt glottal closure. Creaky voice is usually produced with a very low fundamental frequency, and it occurs often at the end of an utterance where the subglottal pressure is low. Creaky voice may also involve a secondary excitation peak due to the opening of the glottis (Gobl and Ní Chasaide, 1992; Blomgren et al., 1998). The temporal excitation pattern of creaky voice can be rather regular if the vocal folds can maintain the low frequency vibration, but often irregular excitation is observed. Creaky voice may also involve a diplophonic excitation pattern (Hedelin and Huber, 1990), where long and short glot-

tal cycles alternate. Creaky voice is characterised with the sensation of perceiving individual pulses due to a low f_0 and strong glottal closure. According to another definition, creaky voice is used to describe a voice quality with very high tension at the larynx, which results in an irregular f_0 and low intensity, whereas vocal fry or *pulse register* is characterised with relaxed vocal folds and and a low f_0 (O'Shaughnessy, 2000).

Finally, in the *falsetto* register, longitudinal tension makes the vocal folds thin, and a small subglottal pressure is used so that the vocal folds vibrate only by the edges of the ligament with only a very small amplitude. The glottis closes only briefly or not at all and the rate of vibration is high, resulting in a steep spectral tilt (Monsen and Engebretson, 1977) and a high pitch, respectively. Falsetto is not normally used in speech, but frequently in singing.

In medical applications, there are additional voice quality classes, such as *hoarse* voice (see other voice quality classes, e.g., in Davis, 1979) which is characterised by strong aperiodicity and aspiration noise, usually due to a pathological condition of the vocal folds.

Different voice qualities and phonation types are used in spoken language for several purposes. Although rare, phonation type has a linguistic function in some languages (see, e.g., Laver and Trudgill, 1979; Gordon and Ladefoged, 2001). Normally, voice quality is used for a paralinguistic function, such as communicating intention, attitude, or affective state (Gobl and Ní Chasaide, 2003). Voice quality or voice source dynamics play an important role in prosody and intonation, signalling prominence or focus (Vainio et al., 2010; Yanushevskaya et al., 2010; Ní Chasaide et al., 2011) in spoken utterances.

Humans also adapt their vocal communication to the acoustic and auditory environment in order to successfully and efficiently deliver a message to a listener without using unnecessary effort. This results in the adjustment of *vocal effort*. For example, in environments with high levels of interfering noise, more effort is required in order to increase the signal-to-noise ratio (SNR) and thereby the intelligibility of speech. This automatic effect, known as the *Lombard effect*, has been widely studied (Lombard, 1911; Junqua, 1993). Depending on the acoustic environment, speech is produced at a different point in the vocal effort continuum (Raitio et al., 2014c), ranging from whispery or soft through normal and finally to Lombard speech and shouting. The change in the vocal effort can be triggered by interfering noise (Summers et al., 1988) or the need to commu-

nicate over a distance (Traunmüller and Eriksson, 2000), but also due to a change in emotional expression (Ishi et al., 2010; Gobl and Ní Chasaide, 2003).

Shouting is the loudest mode of vocal communications and differs from other phonation modes (Rostolland, 1982a,b, 1985). A shout is characterised by an increased sound pressure level (SPL) and a higher f_0 due to the increased subglottal pressure and vocal fold tension, which also results in an emphasis on the higher harmonics (Rostolland, 1982a; Elliott, 2000). In very loud shouting, the phonation also becomes irregular. Shouting is also less articulated and thus less intelligible than Lombard or modal speech (Pickett, 1956; Rostolland, 1985).

2.4 Source-filter theory

Although the speech production mechanism is a single continuous physiological apparatus, the contributions of the glottal excitation and the vocal tract filter to speech can be determined rather independently. This notion has led to the source-filter theory of speech production (Fant, 1960), which considers the *source* and *filter* to be independent from each other. The filter is assumed to be linear time-invariant (LTI), which means that each short-time segment of speech is assumed to have constant parameters without any interaction with the glottal source. These assumptions are clearly simplifications, since the accurate physical description of the generation and propagation of sound in the vocal organs leads to a complex set of differential equations (Rabiner and Schafer, 1978). However, the source-filter theory serves as a useful approach for various speech technology applications, such as speech analysis, coding, and synthesis.

According to the source-filter theory of speech production (Fant, 1960), speech can be uniquely represented as *source* and *filter* characteristics. The primary sound source of voiced speech is the excitation generated by the vibrating vocal folds, called the glottal flow. The spectrum of the quasi-periodic glottal flow signal is rich in harmonics, whose energy declines with increasing frequency on average by 12 dB per octave (Flanagan, 1972a; Kent and Read, 1992). In the case of unvoiced speech, the sound source is assumed to be white noise arising from a constriction at some point in the vocal tract, leading to turbulent air flow. The vocal tract is assumed to be a tube that is closed at the glottis, open at the mouth, and having a varying cross-sectional area. Since the diameter of the tube

is small compared to the wavelength of relevant speech sounds, plane-wave propagation can be assumed. The vocal tract modifies the glottal flow spectrum by creating formants, or antiformants[2] in the case of nasal sounds. Finally, the spectrally modified glottal flow exits through the lips and nostrils and radiates to the surrounding air as a sound pressure wave. This process, transforming the air flow into a pressure wave, is called *lip radiation*. It can be approximated by a time derivative of the flow, which acts as a high-pass filter (Flanagan, 1972a) increasing the magnitude of the spectrum approximately by 6 dB per octave. Assuming an LTI system, where the source and filter are independent, the source-filter theory can be stated in the z domain as

$$S(z) = G(z)V(z)L(z), \tag{2.1}$$

where $S(z)$ is the speech signal, $G(z)$ the glottal excitation, $V(z)$ the vocal tract filter, and $L(z)$ the lip radiation (Fant, 1960; Markel and Gray, 1980). The lip radiation can be approximated by a first-order differentiator

$$L(z) = 1 - \rho z^{-1}, \tag{2.2}$$

where, depending on the definition, ρ is set to a constant close to, but less than, 1. This differentiating operation is commonly combined with the glottal source model $G(z)$ so that the speech production model becomes

$$S(z) = \hat{G}(z)V(z), \tag{2.3}$$

where $\hat{G}(z) = G(z)L(z)$. The vocal tract filter $V(z)$ is usually described as an all-pole linear filter with complex conjugate poles. However, the all-pole model is not ideal for modelling antiresonances, for example, in nasal sounds (Rabiner and Schafer, 1978), which also requires zeros in the transfer function. However, as suggested by Atal and Hanauer (1971), the zeros can be modelled adequately by including more poles in the model. Moreover, zeros are perceptually less important than poles (Malme, 1959; Klatt, 1987), and thus an all-pole filter is usually a reasonable choice. Very often linear prediction (LP) is used for estimating the filter part, capturing the overall spectrum of speech. Therefore, the source spectrum is forced to be approximately flat, ignoring the natural spectral properties of the source. This approach has the benefit of simplicity without trying to specifically estimate the separate contributions of the natural voice source

[2] In the case of nasal sounds, a more complicated tube model needs to be used to model the three cavities: the pharynx, oral cavity, and the nasal cavity.

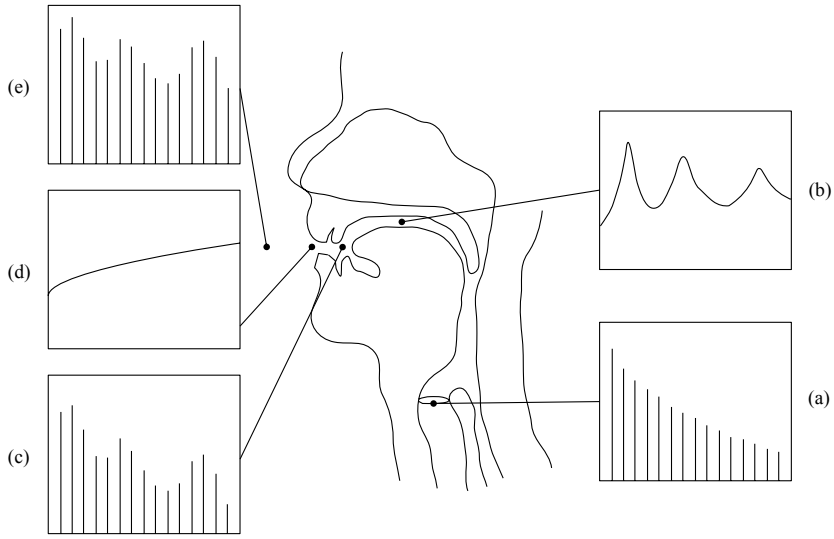


Figure 2.9. Illustration of the source-filter model of speech production. (a) The excitation signal of voiced speech generated by the vibrations of the vocal folds, which produces a rich periodic spectrum, whose energy declines with increasing frequency. (b) The vocal tract modifies the glottal excitation by forming resonances. (c) The spectrum of the signal before radiating from the lips. (d) The radiation of sound from the lips and nostrils to the surrounding air produces an effect that enhances the higher frequencies of the signal. (e) The final spectrum of the speech signal after the lip radiation.

and vocal tract filter, but it also limits the analysis and modelling of natural speech production, for which, for instance, glottal inverse filtering is required (see Section 3.1). The schematic illustration of the source-filter theory is presented in Figure 2.9. The detailed derivation of the theory is presented, for example, in Fant (1960), Flanagan (1972a), and Rabiner and Schafer (1978).

The simplifications made in the source-filter model naturally have some implications that limit the accuracy of the modelled speech. The first issue is the time invariance. For voiced speech sounds such as vowels, the parameters of the quasi-stationary model vary rather slowly and thus the model is sufficient, but for transient sounds, such as stops, the model is not that accurate due to the rapid changes in the excitation signal. However, the model is still perceptually adequate for several applications. The second limitation is the already mentioned incapability of modelling the zeros if all-pole modelling is used. However, by adding more poles to the model transfer function $V(z)$, the lack of zeros can be partially compensated (Atal and Hanauer, 1971). Thirdly, the model does not provide means for modelling, say, voiced fricatives if a binary voiced/unvoiced decision is assumed. Finally, the model does not take into account the in-

teraction between the source and the filter and the nonlinearities therein. The source-filter interaction has been widely studied (see, for example, Rothenberg, 1981; Ananthapadmanabha and Fant, 1982; Ananthapadmanabha, 1984; Fant et al., 1985b; Lin, 1987; Fant and Lin, 1987; Klatt and Klatt, 1990; Teager and Teager, 1990). Effects such as skewing, ripple, and damping of the glottal flow pulse due to the interaction have been observed. Although articulatory synthesis experiments have not provided significant perceptual improvements in speech naturalness when utilising source-filter interaction (Nord et al., 1986; Lin, 1990), it cannot be concluded that the source-tract interaction does not play an important role in speech production and perception. Especially in speech processing applications that aim to modify or generate speech, the source-filter interaction may be crucial for the naturalness of speech. The source-filter interaction has not been used explicitly, for example, in statistical parametric speech synthesis (Zen et al., 2009), which might be one of the reasons for the unnatural speech quality, as is speculated in the studies by Merritt et al. (2014); Henter et al. (2014).

2.5 Characteristics of hearing

The purpose of the hearing system is to transfer information conveyed by the sound pressure waves into meaningful information for further processing in the brain. The first component of the hearing system is the ear, which consists of three regions. The first, the outer ear, consisting of the pinna and the ear canal, funnels sound waves into the ear drum and helps in the localisation of sounds. The resonance of the ear canal boosts the frequencies in the 3–5-kHz range, which aids the perception of sounds at these frequencies (O'Shaughnessy, 2000). The middle ear, consisting of the ear drum and the ossicular bones, serves as an impedance transformer between the air medium and the liquid medium of the inner ear. There is a large boost in sound amplitude at 1 kHz due to the structure of the the middle ear, and it also acts as a low-pass filter above 1 kHz (O'Shaughnessy, 2000). Finally, the ossicular bones transfer the vibrations through the oval window and to the fluid in the cochlea in the inner ear. The moving fluid makes the basilar membrane vibrate at different positions for different frequencies, which results in hair cell movements, causing frequency-dependent neuronal firings in the auditory nerve. The neural information follows a pathway to the brain for further processing,

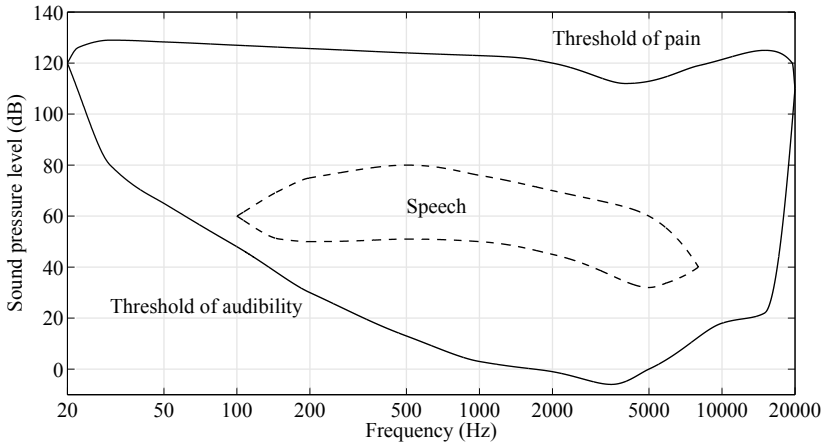


Figure 2.10. Approximate hearing range in terms of frequency and sound pressure level (SPL), and the most important area for speech perception thereof (adapted from Rossing et al., 2002; O'Shaughnessy, 2000). SPL is defined $L_p = 20\log_{10}(p_{\mathrm{rms}}/p_0)$, where p_{rms} is the root mean square sound pressure, and p_0 is the reference sound pressure, both measured in Pa. Here $p_0 = 20\,\mu\mathrm{Pa}$.

leading finally to the auditory cortex.

Human hearing ranges approximately from 20 Hz to 20 000 Hz (Rossing et al., 2002), but since hearing most probably has evolved in parallel with speech production, hearing is most sensitive at those frequencies relevant to speech communication, that is, 200–5600 Hz (O'Shaughnessy, 2000). The human hearing range with respect to the sound pressure level (SPL) and frequency is illustrated in Figure 2.10. The range commonly used in speech is also depicted, although the range depends slightly on the speaker and the speaking style. For example, in shouting or whispering, the SPL may be higher or lower than the depicted area, respectively.

The perception of sounds depends both on the frequency and the SPL, but in a rather complex and nonlinear manner due to the behaviour of the cochlea and the neural processing. The perception of pitch with respect to the actual frequency is nonlinear due to the characteristics of the basilar membrane; low frequencies make the basilar membrane vibrate over a much wider area than high frequencies do. Therefore, the perceived pitch with respect to frequency is logarithmic, and it can be approximated using, for example, the mel-scale (Stevens et al., 1937)

$$m = 2595\log_{10}(1 + f/700), \tag{2.4}$$

where f is the frequency in Hz and m is the perceived pitch in mel (O'Shaughnessy, 2000). For many applications, it is thus convenient to describe the frequency-related quantities, such as pitch or formant fre-

quencies, with perceptually weighted auditory scales instead of linear frequency.

The basilar membrane consists of tonotopically organised hair cells (O'Shaughnessy, 2000), which form the so called *critical bands* (Fletcher, 1938a,b). Critical bands are frequency bandwidths inside which the perception of two tones interfere with each other. Critical bands can be described by, for example, the Bark scale (Zwicker, 1961) or the equivalent rectangular bandwidth (ERB) scale (Moore and Glasberg, 1983). The perception of sound is highly dependent on the relative energies in different critical bands as well as their distribution in time. This has effects on loudness perception and gives rise to effects such as *masking* (O'Shaughnessy, 2000) in time and frequency, and *interference* (Moore, 2002). Due to the masking effect in frequency (among other effects), the spectral peaks are more easily perceived than spectral valleys (Malme, 1959), which is an especially important feature in speech perception, where formants mostly define the identity of different speech sounds while spectral valleys mostly affect the amplitudes of the nearby formants (Klatt, 1987).

A common conception has long been that human hearing is not sensitive to phase due to early studies on the topic (Ohm, 1843; von Helmholtz, 1863). Although phase information is perceptually not as important as spectral information, subsequent studies (see, e.g., Schroeder, 1959; de Boer, 1961; Plomp and Steeneken, 1969; Bilsen, 1973; Carlson et al., 1979; Schroeder and Strube, 1986; Patterson, 1987; Moore and Glasberg, 1989; Moore, 2002) show that phase plays a perceptually important role in certain signals, such as in speech (Carlson et al., 1979; Schroeder and Strube, 1986; Pobloth and Kleijn, 1999; Laitinen et al., 2013; Drugman and Raitio, 2014). The mechanism of phase perception is based on the phase-locking of the neuronal firing in the cochlea. If a single sinusoid is presented to the hearing system, the neuronal firing rate of the hair cells shows a pulse for each period of the sinusoid corresponding to the specific value of the phase of the sinusoid. Thus, a sound is perceived differently if the neuronal firing between different frequencies are synchronous or asynchronous.

However, phase perception is only limited to signals with a low repetition rate (i.e., f_0 in speech) (Patterson, 1987). Since the electrical pulses fired by the auditory nerves have a finite duration, the phase-locking effect is lost in the high frequencies (Joris et al., 1994). The pulse duration

is 0.5–1.0 ms, depending on the neuron type and frequency (Joris et al., 1994), and the temporal accuracy of hearing is approximately 1–2 ms with impulsive signals (Moore, 1982). Various experiments on phase perception have been conducted. Patterson (1987) suggests that humans are phase deaf for signals with repetition rate approximately above 400 Hz but not below 200 Hz. In Laitinen et al. (2013), in-phase and random-phase periodic signals with repetition rates from 50 Hz to 1600 Hz were used as test stimuli, and the difference due to phase was found to gradually decrease with the repetition rate. Beyond 800 Hz, test subjects could not reliably tell the difference between the two signals based on the phase differences.

Phase perception has a complex dependency on repetition rate, signal intensity, bandwidth, and phase characteristics (Patterson, 1987; Laitinen et al., 2013). Humans are most sensitive to harmonic complex signals in which the phases of the harmonics have certain fixed relations, such as in speech, trumpet, and trombone (Laitinen et al., 2013). All these signals originate from a physiological sound source, that is, the fluctuations of the vocal folds or lips. Phase perception has important implications for speech technology, as the naturalness of most men's voices and many women's voices, depending on the f_0 of the speaker, is dependent on the phase relation of the harmonic components (Patterson, 1987).

2.6 Speech perception

From physiological perspective, speech perception begins as the sound-pressure waves of speech arrive at the ear and create neuronal firing in the cochlea. The neural information is further processed to extract and perceive acoustic cues for speech, and to classify this information into meaningful entities based on the learnt language. Finally, the speech information is used for higher level speech and language processing, for example, using Wernicke's and Broca's areas in the brain.

In the auditory system, frequencies from 200 Hz to 5600 Hz contribute most to the perception of speech signals (O'Shaughnessy, 2000). Hearing is most sensitive in this frequency range, which is where most of the speech energy is also concentrated. A speech signal can be band-limited to about 10 kHz with only minor effects on its perception (Paliwal and Kleijn, 1995), and rather intelligible speech can be obtained with a much narrower bandwidth, such as is used in the traditional telephone band

of 300–3400 Hz. However, the bandwidth corresponding approximately to the full hearing range is required for reproducing completely natural sounding speech. In speech synthesis, a sampling rate of 16 kHz is commonly utilised for simplicity, which enables reproducing frequencies up to 8000 Hz. This yields rather good speech quality, although some higher frequencies are lost. Also higher sampling rates can be utilised, such as 44.1 kHz or 48 kHz, which enable the reproduction of all frequencies in human hearing.

From the human perspective, speech perception can be described using subjective terms such as intelligibility, naturalness, expressivity, and speaker identity. All of these are ultimately defined by the acoustic properties of the speech signal, but different acoustic cues have varying degrees of relevance, depending on the measured quantity. For example, natural speech contains multiple redundant acoustic cues for the perception of phonemes, but the irrelevant cues may have other functions, for instance in terms of recognising the speaker. Most sounds are perceived on a continuous scale as the acoustic cues change, but certain stimuli, especially many speech sounds, are perceived categorically, that is, the ability to discriminate two sounds depends on labelling them as linguistically different. This has important implications in speech understanding, where speech sounds with varying acoustic cues due to, say, context and speaker, can be perceived to have the same linguistic meaning. Most current TTS systems are able to deliver rather intelligent speech by reproducing the required phonetic information but may not succeed in reproducing all other cues that contribute to naturalness and contextual aspects of speech. In the following, the most salient cues for recognising different speech sounds are described, after which other perceptually relevant cues in speech are discussed.

Vowels are perceived when the sound is periodic and it has sufficient energy and duration and a strong formant structure. Vowel perception is relatively simple since the positions of the three first formants define the vowel identity (Hillenbrand and Gayvert, 1993; Pickett, 1999). The higher formants remain rather constant regardless of changes in articulation. The perception of vowels is very sensitive to the formant location in frequency and also to the formant amplitude, but less to its bandwidth. The just-noticeable difference (JND) for the first and second formant frequencies have been measured to be 3–5% of the centre frequency (Flanagan, 1972a). The formant amplitude JND is estimated to be 1.5 dB and

3 dB for the first and second formants, respectively (Flanagan, 1972a). However, changes in formant bandwidth (−3 dB) of the order of 20–40% have been found to be just noticeable (Flanagan, 1972a).

The perception of consonants is more complex, since the categorisation is based on multiple cues, such as spectrum, amplitude, voicing, and duration as well as on the interaction with surrounding phones (O'Shaughnessy, 2000). Different consonants are often distinguished from each other using cues such as noise burst frequency distribution, voice onset time (VOT) (Lisker and Abramson, 1964), formant transitions before and after the consonant, and duration of a closure.

However, several factors may alter the acoustic properties of the prototypical speech sounds, such as coarticulation, context, speaker, and speaking style. In coarticulation, adjacent speech sounds become more similar to each other due to the physical constraints of the articulatory movements. Changing from one sound to another is performed in an efficient manner so that the trajectories of the articulators are smoother. Thus, the articulators might not be in the final position of each phoneme, but somewhere between that and that of the adjacent phoneme. This effect is also called *reduction*, *undershooting*, or *assimilation*. Also, the hypothetical boundaries between different (isolated) phonemes appear as a continuous change from one phone to another, which makes it impossible to exactly define a discrete point where one phoneme ends and the next begins. This is illustrated in Figure 2.11, which shows a speech signal, its spectrogram, and the approximate boundaries between different phonemes. Coarticulation is important for producing smooth and connected speech, and it helps in auditory stream integration and continuity (Cole and Scott, 1974). However, in speech recognition, defining the boundaries of the underlying phonemes based on the acoustic cues becomes difficult. In speech synthesis, if coarticulation is not modelled, different speech sounds, such as vowels and fricatives, are heard as two separate sound streams instead of connected speech (O'Shaughnessy, 2000). Coarticulation is increased in conversational, expressive, casual (hypo-articulated) speech, and in speech with an increased speaking rate, while in formal, read-aloud, and hyper-articulated or clear speech, coarticulation is decreased (Lindblom, 1983). Coarticulation varies also in terms of speakers and their motivation, emotion (Beller et al., 2008), and relation to the listener.

The rhythm, stress, and intonation of speech, collectively called as *prosody*, is used to help the listener understand the message by pointing

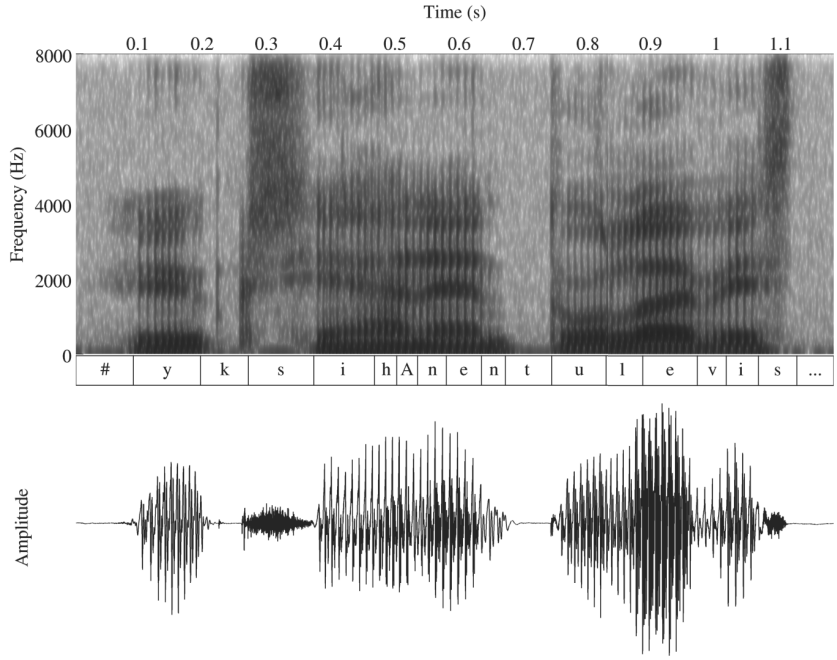


Figure 2.11. Spectrogram (upper) and the speech signal (lower) of a Finnish utterance "Yksi hänen tulevis…". The aligned monophone transcription is shown below the spectrogram. The hash (#) symbol denotes a silence.

out important words and creating logical breaks in the speech flow, thus segmenting and highlighting speech. Prosody is also used to convey additional information, such as the utterance type (statement, question, command), irony or sarcasm, emphasis, contrast, focus, and emotional state of the speaker. For example, emotions have a major effect on speech prosody (Scherer, 2003; Laver, 1980), introducing changes to various speech features at both the segmental and the suprasegmental level. Emotions affect the intonation, duration, intensity, speaking rate, and articulation as well as the phonation of speech. In noisy environments, prosody can also be used as a continuity guide to follow a specific speaker (Brokx and Nooteboom, 1981). Prosody extends beyond phonemes to syllables, words, phrases, sentences, and even to longer term structures. The acoustic cues conveying this information are mainly duration, intensity and f_0, but also voice quality (Sluijter et al., 1997; Campbell and Mokhtari, 2003; Vainio et al., 2010; Yanushevskaya et al., 2010; Ní Chasaide et al., 2011) and the degree of articulation (Pfitzinger, 2006; Beller et al., 2008) have been shown to have prosodic functions. Almost all prosodic cues originate from the voice source while only minor changes in prosody are conveyed by articulation. Rhythm, intonation, stress, and prominence, for example,

are subclasses of prosody that relate to the changes in different acoustic cues and their combinations. Although intonation, defined by pitch, is perceptually very important, it seems that the fine variations in pitch are not easily perceived (Rosenberg et al., 1971). The most salient pitch cues are large rises and falls, with more emphasis on rises ('t Hart, 1974) and high-energy (vowel) parts of speech (Léon and Martin, 1972). Also the perception of pitch slope is rather accurate (Klatt, 1973), which indicates its perceptual importance in prosody.

Finally, speakers vary substantially from each other in terms of gender, size, age, and other individual differences. The largest difference between speakers is due to gender. Male speech is generally different from female speech due to the differences in the size and shape of the vocal folds and the vocal tract. Children are also physically very different from adults, which results in a different type of speech. In addition, each speaker has a personal voice based on the properties of the vocal organs. Finally, the individual differences in speaking style, such as language, accent, speech rate, and dialect, affect the use of the speech production organs, and thus each individual has a specific character in their speech. In speech technology, these personal traits are dealt with differently depending on the application. In speech recognition, the acoustic effects of personal traits should be normalised or removed in order to achieve good recognition accuracy for any speaker (Leggetter and Woodland, 1995). In contrast, in speech synthesis, it is often desired to reproduce the personal traits in order to make the speech sound natural, and usually as close as possible to a specific voice (Yamagishi et al., 2009a). Speaker traits can be also used for speaker recognition (Kinnunen and Li, 2010), identification, and verification (Kinnunen et al., 2006).

2.7 Summary

This section described the human speech production mechanism, elaborating on its two distinct parts, the glottal excitation and the vocal tract, and their function in speech production. The properties and classification of different speech sounds in spoken language were briefly covered, after which voice quality and its functions were discussed. The simplified source-filter theory of speech production, widely used in speech technology, was also described. Finally, the hearing mechanism and its implications to speech perception were discussed, and the acoustic cues for speech

perception and contextual effects on speech were presented.

3. Voice source estimation and parameterization

Extracting information originating from the voice source is used virtually in all speech technology applications. The most common speech features that originate from the voice source are speech energy and pitch. However, since the voice source conveys a lot of perceptually relevant information, it is often useful to look for more detailed characterisation. The estimation of the voice source signal from speech, although not trivial, provides a rich source of information and allows the detailed analysis and modelling of the speech signal. This section describes the estimation of the voice source signal through glottal inverse filtering and presents methods and applications on how the voice source information can be further utilised.

3.1 Glottal inverse filtering

Glottal inverse filtering (GIF) is a technique for estimating the glottal volume-velocity waveform from a recorded speech signal. This is performed by cancelling the effects of the vocal tract and lip radiation from the speech signal, thus revealing the time-domain waveform of the glottal source. Since the lip-radiation effect can be approximated with a fixed first-order differentiator (Flanagan, 1972a), the challenge of GIF is in the estimation of the vocal tract filter. GIF is a difficult inverse problem since a speech signal can be decomposed into the two components, glottal flow signal and vocal tract filter, in infinitely many ways. Thus, a priori information about the characteristics of the glottal flow and vocal tract filter must be utilised for successful glottal flow estimation.

According to the source-filter theory of speech production (Fant, 1960), speech can be defined as $S(z) = G(z)V(z)L(z)$, where $S(z)$, $G(z)$, $V(z)$, and $L(z)$ denote the z-transforms of the speech signal, glottal flow signal, vocal

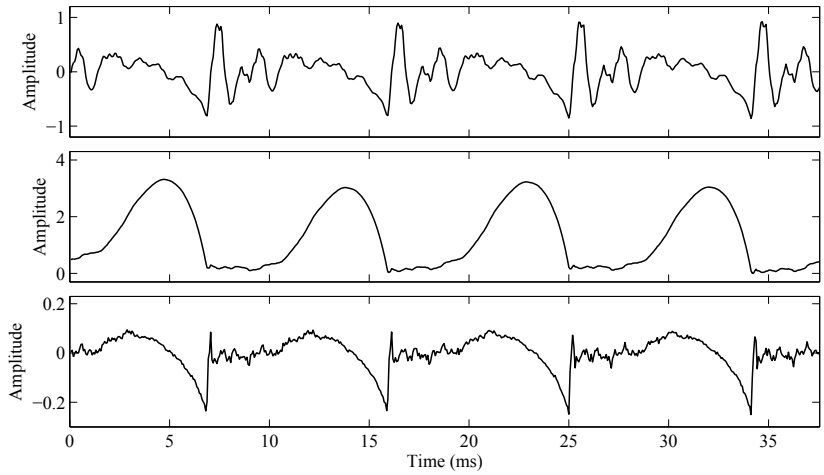


Figure 3.1. Illustration of a speech signal of male vowel [a] (top) and corresponding glottal flow (middle) and differentiated glottal flow (bottom) estimates.

tract resonances, and the lip-radiation effect, respectively. Conceptually GIF corresponds to obtaining the glottal volume velocity $G(z)$ from the equation

$$G(z) = \frac{S(z)}{V(z)L(z)}, \tag{3.1}$$

where $L(z)$ is a fixed differentiator. Thus, the parameters of $V(z)$ need to be estimated to solve $G(z)$.

In the GIF process, the two components $G(z)$ and $V(z)$ should correspond as closely as possible to the physiological phenomena at the glottis and the vocal tract, respectively. This is the main difference between glottal inverse filtering and conventional inverse filtering using, for example, LP, where the filter captures the overall spectral characteristics of the speech signal and the source signal, namely the LP residual, is spectrally white. In contrast to the LP inverse filtering, the voice source signal of GIF is allowed to have a varying spectrum depending, for example, on the phonation type.

An illustration of GIF of the vowel [a] uttered by a male speaker is shown in Figure 3.1. The figure shows the original speech waveform, the estimated glottal flow, and the differentiated glottal flow estimate waveforms.

3.1.1 Glottal inverse filtering methods

The first study on GIF was published by Miller (1959), where a manually tuned analog circuit was used to cancel the first formant of a speech sig-

nal. Since then, GIF has developed from using analog circuitry to digital signal processing (DSP), and from manually tuned to completely automatic methods. Modern GIF methods can be divided roughly into three categories: closed-phase (CP) methods, phase-based methods, and iterative methods. In the following, a short review is given on GIF methods that are capable of automatically estimating the glottal flow from a speech signal recorded outside the lips.

With the introduction of LP (Makhoul, 1975) in speech processing, the automatic computation of the vocal tract filter became an attractive alternative to older, manually tuned techniques. One of the oldest automatic GIF methods utilising LP is the closed-phase covariance method (Strube, 1974; Wong et al., 1979), which uses LP with the covariance criterion (Rabiner and Schafer, 1978) for estimating the vocal tract all-pole response during the CP of the glottal excitation. In theory, the CP is optimal for estimating the vocal tract transfer function since the effect of the glottal excitation to the vocal tract filter response is minimal when the glottis is closed. Indeed, CP analysis yields accurate estimates of the glottal flow for modal speech with a well-defined CP (Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986). However, the method is very sensitive to the estimation of the CP position since even a slight error in the estimated position may significantly distort the vocal tract estimate (Larar et al., 1985; Riegelsberger and Krishnamurthy, 1993; Yegnanarayana and Veldhuis, 1998). This problem can be partly alleviated by using two-channel analysis (Krishnamurthy and Childers, 1986), where the CP boundaries are estimated from the electroglottography (EGG) signal instead of the acoustic speech signal. However, the EGG signal may not be available for the given speech data, and the recording of the EGG signal requires special arrangements and equipment. Another downside of the two-channel analysis is the undetermined (and often varying) acoustic delay between the EGG signal and the microphone. The estimation accuracy of the CP method also remains poor with high-pitched or breathy speech where the CP is either short or absent due to incomplete glottal closure.

Even before the time LP was used in speech processing, GIF experiments utilising DSP were conducted by using homomorphic analysis of speech (Oppenheim and Schafer, 1968). In this approach, speech is transformed into additive components using cepstral analysis. This technique was later refined into a method called zeros of the z transform (ZZT)

(Bozkurt et al., 2005, 2007; Sturmel et al., 2007), or if formulated otherwise, the method is called complex-cepstrum-based decomposition (CCD) (Drugman et al., 2009a). The methods are based on decomposing speech into source and filter characteristics based on the mixed-phase character of the speech signal, which is composed of the maximum phase (anti-causal) glottal excitation and the minimum phase (causal) vocal tract filter. The ZZT method utilises the z transform of a speech frame (Bozkurt et al., 2005), and separates the roots based on their position in the unit circle. Roots that are outside the unit circle correspond to the anti-causal open phase of the glottal source, and the roots inside the unit circle correspond to the causal vocal tract filter and the return phase of the glottal source. The CCD method utilises complex cepstrum to perform the separation, which is computationally more efficient than the ZZT decomposition. Since both ZZT and CCD require the estimation of the glottal closure instants (GCIs), these methods are prone to errors in GCI estimation. A major drawback of these methods is that there is no separation of the return phase from the vocal tract filter, which is required in many applications, especially in those that require the estimation of the complete glottal flow waveform, such as speech synthesis.

Several GIF methods based on iterative glottal flow estimation have been proposed. Matausek and Batalov (1980) have proposed estimating the spectral contribution of the glottal flow by using simple low-order autoregressive (AR) modelling. In their work, the final glottal flow estimate is achieved by filtering an impulse train through the AR model. A more elaborated method called the iterative adaptive inverse filtering (IAIF) was proposed by Alku (1992), in which low and high-order LP or discrete all-pole modelling (DAP) (El-Jaroudi and Makhoul, 1991) is used successively in order to estimate the glottal flow and vocal tract contributions, respectively. IAIF has been used and evaluated in various experiments (see, e.g., Alku et al., 2006a,b; Drugman et al., 2012a). IAIF has been shown to yield rather robust estimates of the glottal flow. However, IAIF is prone to the biasing effect of the excitation with high-pitched speech, especially for vowels with a low first formant frequency.

GIF requires the accurate estimation of the vocal tract spectrum in order to cancel the formant frequencies from the speech signal. However, the periodic glottal excitation makes this estimation rather hard, since its harmonic structure biases the formant estimates towards harmonic peaks. This effect is especially prominent with high-pitched speech, where

the harmonics are sparse. The effect of the excitation on the glottal flow estimate can be reduced, for instance, by using CP analysis that estimates the spectrum from those parts of the speech signal that are not corrupted by the excitation, or using spectral estimation methods that are more robust to the biasing effect of the harmonics, such as DAP (El-Jaroudi and Makhoul, 1991) or weighted linear prediction (WLP) (Ma et al., 1993). Alternatively, the bias can be diminished by jointly estimating the vocal tract filter and a glottal flow model. Such approaches were utilised, for example, in the studies by Milenkovic (1986) and Fujisaki and Ljungqvist (1987), where autoregressive moving average (ARMA) modelling is used to allow the modelling of both spectral peaks (poles) and dips (zeros). In these methods, the zeros are assumed to stem from the voice source, and the Fujisaki–Ljungqvist (Fujisaki and Ljungqvist, 1986) glottal flow model is assumed as an input. Similar methods were proposed in the work by Ding et al. (1997) and Kasuya et al. (1999), in which ARX (autoregressive with exogenous input) modelling is utilised using the Rosenberg–Klatt (RK) glottal flow model (Rosenberg, 1971; Klatt, 1980; Klatt and Klatt, 1990) as an input. Fröhlich et al. (2001) have also proposed a method that utilises ARX modelling through DAP (El-Jaroudi and Makhoul, 1991) for spectrum estimation and the Liljencrants–Fant (LF) glottal flow model (Fant et al., 1985a; Fant, 1995) as an input. Similarly, the LF glottal flow model is also used as an input within ARX modelling in the work by Fu and Murphy (2003, 2006).

More extensive reviews on GIF can be found, for example, in the studies by Walker and Murphy (2005), Alku (2011), and Drugman et al. (2014). This thesis presents two new GIF methods that aim to reduce the biasing effect of the excitation. First, a GIF method using joint estimation of a simple glottal flow model and the vocal tract spectrum through the Markov chain Monte Carlo (Gilks et al., 1996; Hastings, 1970; Gamerman, 1997; Smith and Roberts, 1993; Tierney, 1994; Roberts and Smith, 1994) algorithm is presented in Publication I (Auvinen et al., 2012, 2014). Next, in Publication II, another new GIF method is introduced that uses the idea of CP analysis, but in this method the analysis is performed over multiple fundamental periods using WLP (Ma et al., 1993) and a specific attenuation function aligned with the glottal closure instants in order to reduce the effect of the excitation on the spectrum estimation (Airaksinen et al., 2014).

3.1.2 Applications of glottal inverse filtering

GIF finds applications in several areas of speech technology and research. First of all, it has an important role in the research of voice communication. Information on the glottal flow characteristics, obtained through GIF, have been used to study, for example, general speech production, phonation type, voice quality, vocal emotions, intensity regulation, prosodic features, singing voice, ageing, and source-tract interaction. Secondly, glottal flow estimation finds applications in medical research, such as in the analysis of pathological voices and assessing vocal loading. Finally, GIF is used in various speech technology applications. Speech synthesis is probably the oldest and the most obvious application area, where artificial glottal source models have been used to excite the formant based vocal tract filter to create synthetic speech (see, e.g., Klatt, 1987; Carlson et al., 1989; Pinto et al., 1989; de Veth et al., 1990; Klatt and Klatt, 1990; Carlson et al., 1991; Karlsson, 1991, 1992; Fant, 1993; Childers and Hu, 1994; Childers, 1995; Childers and Ahn, 1995). In addition to these methods, which mostly utilised the Liljencrants–Fant (LF) glottal flow model (Fant et al., 1985a; Fant, 1995), glottal flow pulses estimated from natural speech have been used to create the excitation in speech synthesis (Holmes, 1973; Matsui et al., 1991; Karjalainen et al., 1998; Fries, 1994; Alku et al., 1999). Although the use of glottal flow excitation in speech synthesis has become less popular due to the emergence of unit selection synthesis, GIF has recently found novel applications in vocoding for statistical parametric speech synthesis (Raitio, 2008; Raitio et al., 2008, 2011c,a). For example, GIF is used for speech synthesis in Publications III, IV, V, VII, and VIII. Other areas that utilise GIF include, for example, voice modification and conversion, speaker identification, dialect identification, emotion and speaking style classification, speech coding, and aid for phonetic segmentation.

3.2 Glottal flow parameterization

Usually the aim of GIF is not only to reveal the underlying glottal flow waveform, but to express the information obtained from the estimated voice source signal in a meaningful manner. This calls for the parameterization of the glottal flow that represents the most important features of the computed waveform in a compressed numerical form. Glottal flow pa-

rameterization plays an important role in voice production research and in various speech technology applications. A large number of different parameterization methods have been developed, most of which are discussed in this section. The parameters are divided into three categories: time-domain parameters, frequency-domain parameters, and parameterization by model fitting. However, first the issue of glottal closure instant detection is discussed, which is important both for many GIF methods as well as for several glottal parameterization methods.

3.2.1 Glottal closure instant detection

Glottal closure instant (GCI) detection (Naylor et al., 2007) aims to find the exact time positions of the GCIs from a speech signal. This is the same task as epoch detection (Ananthapadmanabha and Yegnanarayana, 1975; Smits and Yegnanarayana, 1995; Murty and Yegnanarayana, 2008), since epoch is defined as the point of maximum discontinuity in the derivative of the glottal waveform (Flanagan, 1972a). GCI detection is important for various applications (see a review by Yegnanarayana and Gangashetty, 2011), such as speech analysis, modification and transformation (Moulines and Charpentier, 1990; Rao and Yegnanarayana, 2006; Agiomyrgiannakis and Rosec, 2009), GIF (Wong et al., 1979; Alku, 1992; Bozkurt and Dutoit, 2003; Bozkurt et al., 2005; Drugman et al., 2009a), data-driven voice source modelling (Thomas et al., 2009; Gudnason et al., 2012), and speech synthesis (Stylianou, 2001; Raitio et al., 2008, 2011c,a; Drugman et al., 2009b; Drugman and Dutoit, 2012).

Although GCI detection can be performed without actual GIF, most methods aim to remove the contribution of the vocal tract system from the speech signal, either by GIF or deriving an LP residual as a preprocessing step for GCI detection. GCI detection is discussed here for two reasons: it is closely related to GIF, and it is crucial in many methods concerning further glottal flow processing.

With the introduction of advanced data-driven speech processing methods and applications, there is an increasing demand for automatic and robust GCI detection methods. Recently, various GCI detection methods have been proposed (see reviews, for example, by Drugman, 2011; Drugman et al., 2012c; Kane, 2012). Also, GCI detection methods for use in adverse conditions (Drugman and Dutoit, 2009) or to improve performance with various voice qualities (Cabral et al., 2011a; Kane and Gobl, 2013) have been studied. GCI detection algorithms can be based

on various different methods, like the Hilbert envelope of the LP residual (Ananthapadmanabha and Yegnanarayana, 1975, 1979; Cheng and O'Shaughnessy, 1989; Rao et al., 2007), the Frobenius norm (Ma et al., 1994), the wavelet transform to find discontinuities in the speech signal (Kadambe and Bourdreaux-Batels, 1992; Tuan and d'Alessandro, 1999; Sturmel et al., 2009), weighted nonlinear prediction (Schnell, 2007), detecting abrupt changes in the short-term spectral characteristics (Moulines and Di Francesco, 1990), and the group delay function (Smits and Yegnanarayana, 1995). However, maybe the most widely used methods today are the DYPSA algorithm (Kounoudes et al., 2002; Naylor et al., 2007) that uses the phase-slope function and dynamic programming (DP); ESPS (Talkin, 1989, 1995) that utilises detecting the maxima in the short-term energy (STE) normalised LP residual and DP; YAGA (Thomas et al., 2012) that combines several methods, such as wavelet analysis, the group delay function and DP, zero-frequency filtering (ZFF) (Murty and Yegnanarayana, 2008) that uses a zero-frequency resonator and mean subtraction to reveal the epochs; and SEDREAMS (Drugman and Dutoit, 2009; Drugman et al., 2012c) that uses the mean-based signal to estimate approximate regions of GCIs and then detects the maxima in the LP residual. To improve detecting the GCIs from speech with varying voice quality, Kane (2012) and Kane and Gobl (2013) have proposed a method called SE-VQ that extends the SEDREAMS algorithm by, for example, adding DP and post-processing. In the work of the present author (Raitio et al., 2011a, 2013b, 2014a,b), a method close to the SEDREAMS algorithm is used, but the determination of the approximate time windows for the GCI search is performed differently. Instead of using the mean-based signal, a global minimum of the glottal flow derivative is first sought within a frame, which is assumed to be the most significant GCI. From that time instant, the preceding and following GCIs are sought within predefined time windows at fundamental-period intervals.

Often the ground truth GCIs are estimated from the EGG signal. EGG is free from the resonances of the vocal tract and aspiration noise, and thus the estimation of GCIs from EGG is more reliable. Alternatively, the voice source can be characterised using an accelerometer signal (Mehta et al., 2012). However, without special recording equipment for EGG or accelerometer signal, the only practical (non-invasive) way of estimating GCIs is to utilise the acoustic speech signal. In addition to GCIs, sometimes the goal is to estimate the glottal opening instants (GOIs), such

as in the work by Brookes and Loke (1999), Bouzid and Ellouze (2004), Drugman and Dutoit (2009), and Thomas et al. (2012).

It is also important to note that many glottal flow estimation and processing methods require that the analysed speech signal has the correct polarity. The polarity has relevance due to the asymmetry of the glottal excitation. In the correct polarity, which means that the speech signal is not inverted due to, say, the recording equipment, the glottal closure instant in the differentiated glottal flow waveform shows a negative peak. There are several methods for polarity detection (see, for example Ding and Campbell, 1998; Saratxaga et al., 2009; Drugman and Dutoit, 2013; Drugman, 2013). For a review and comparison of speech polarity detection algorithms, see the extensive study by Drugman and Dutoit (2014).

3.2.2 Time-domain parameters

Since glottal flow is a quasi-periodic signal, it is straightforward to define simple time-based quantities, such as the GCI, GOI, and the instant of maximum glottal flow. Based on these instants, several quotients can be defined, such as the open quotient (OQ) (Timcke et al., 1958), speed quotient (SQ) (Timcke et al., 1958), and the closing quotient (ClQ) (Monsen and Engebretson, 1977). Sometimes the closed quotient (CQ) is used instead of OQ (Iwarsson et al., 1998; Sundberg et al., 1999a,b). Also the return quotient (RQ) can be used for characterising the return phase. The quantities are defined as

$$\begin{aligned} \mathrm{OQ} &= (t_\mathrm{o} + t_\mathrm{cl})/T_0, \\ \mathrm{SQ} &= t_\mathrm{o}/t_\mathrm{cl}, \\ \mathrm{ClQ} &= t_\mathrm{cl}/T_0, \\ \mathrm{CQ} &= t_\mathrm{c}/T_0, \\ \mathrm{RQ} &= t_\mathrm{ret}/T_0, \end{aligned} \tag{3.2}$$

and their parameters are illustrated in Figure 3.2. These quantities have been used in various studies related to, for example, gender, age, pitch, loudness (Holmberg et al., 1988; Sulter and Wit, 1996), assessment of vocal disorders (Hillman et al., 1989), measuring vocal loading (Vilkman et al., 1997), speech perception (Childers and Lee, 1991), and singing voice (Sundberg et al., 1993).

Often the estimation of specific time instants in the glottal flow is rather ambiguous due to the formant ripple originating from incomplete cancelling of formants by the inverse filter or due to the noise present in the

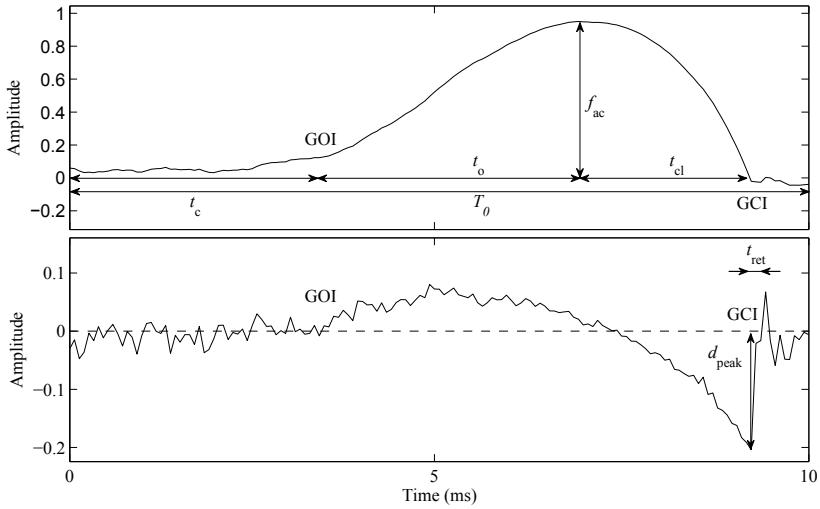


Figure 3.2. Time and amplitude characteristics of the glottal flow (upper graph) and its derivative (lower graph).

waveform, which may originate either from the voice source itself (aspiration noise) or from the recording environment (Alku, 2011). Also, the gradual opening of the glottis makes estimating the time instant rather difficult. Therefore, parameters based on the approximate time instants, defined when a glottal flow crosses a certain amplitude ratio (e.g., 50%) compared to the maximum amplitude, have been proposed (Dromey et al., 1992; Sapienza et al., 1998). Maybe the most widely used such parameter is the quasi-open quotient (QOQ) (Hacki, 1989), which is defined as the time during which the glottal flow amplitude is higher than 50% of the maximum level divided by the pitch period. In order to alleviate the difficulty of measuring the time-based features, the glottal flow can be also characterised by measuring the amplitude features from both glottal flow and its derivative. One such parameter is the normalised amplitude quotient (NAQ) proposed by Alku et al. (2002), which is defined as the ratio between the maximum amplitude of the glottal flow and the negative peak amplitude of the glottal flow derivative

$$\mathrm{NAQ} = f_{\mathrm{ac}}/(d_{\mathrm{peak}} T_0). \tag{3.3}$$

The ratio of these two values can be shown to be a time-domain quantity. Since the two values are the maximum and minimum of the glottal flow and its derivative, respectively, they are easy to extract. NAQ has a close relation to ClQ and voice source parameter R_{d} (Fant et al., 1994; Fant, 1995, 1997) in the LF glottal flow model (Fant et al., 1985a; Fant, 1995), but it has been shown, for instance, that NAQ is more robust to

noise than ClQ (Alku et al., 2002; Bäckström et al., 2002). Airaksinen and Alku (2014) also proposed a phase plane symmetry (PPS) parameter, which is based on the properties of the phase-plane plot (Edwards and Angus, 1996). The PPS parameter was shown to perform similarly to NAQ.

In addition to the time-domain parameters that characterise the properties of the glottal waveform based on a few critical time instants, such as GCI or instant of maximum flow, data-driven voice source waveform modelling approaches have been proposed. These methods are based on segmenting the voice source waveform into individual cycles, normalising each cycle in scale and amplitude (depending on pitch and energy, respectively), and then using machine learning techniques for constructing a voice source model. In Thomas et al. (2009), a set of prototype waveform classes is derived that can be used for the analysis and synthesis of an unknown utterance based on mel-frequency cepstrum coefficients and Gaussian mixture modelling. Gudnason et al. (2009, 2012) use similar approach, but they utilise principal component analysis (PCA) for the data-driven voice source modelling. Various PCA-based voice source modelling techniques have been utilised for statistical parametric speech synthesis, (Drugman et al., 2009b; Drugman and Dutoit, 2012; Sung et al., 2010; Raitio et al., 2013b, 2014c; Drugman and Raitio, 2014). Also a deep neural network (DNN) based voice source modelling method was proposed in Raitio et al. (2014a) and utilised for synthesising various voice qualities in Raitio et al. (2014b).

3.2.3 Glottal flow models

Instead of characterising the estimated voice source signal with individual amplitude or time-based parameters, it is possible to fit a mathematical glottal flow model to the estimated waveform or its derivative. In this approach, a glottal flow model is first selected, after which the model parameters are fitted to the estimated glottal flow so that the error between the two waveforms, the artificial model and the natural one, is minimised. The fitting can be performed in several ways depending on the glottal flow model (see, e.g., Strik et al., 1993; Strik, 1998; Airas, 2008; Kane, 2012). There are several artificial glottal flow models of which perhaps the simplest one is the third order polynomial, the Klatt model (Klatt, 1980),

which is defined as

$$g(t) = \begin{cases} at^2 + bt^3, & 0 \leq t \leq \mathrm{OQ} \cdot T_0, \\ 0, & T_0 \cdot \mathrm{OQ} < t \leq T_0, \end{cases} \tag{3.4}$$

where t is time and T_0 is the length of the glottal cycle. Since OQ defines the numerical values for a and b, the model effectively has only two parameters, OQ and T_0. However, the return phase in the Klatt model cannot be modelled, which limits the use of this model in many applications. A more complex and the most widely used glottal flow model is the LF model proposed by Fant et al. (1985a) and Fant (1995), in which the glottal flow derivative consists of two separate waveform segments. The first segment models the derivative waveform from one glottal opening to the following glottal closure using an exponentially decaying sinusoid. The second segment models the closure of the vocal folds after the abrupt flow termination with a set of exponential terms, causing the flow derivative to return to zero with a specific time constant. The LF model is uniquely defined by 4 parameters:

$$g(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega_\mathrm{g} t), & t < t_\mathrm{e}, \\ -E_\mathrm{e}/(\epsilon t_\mathrm{a}) \left(e^{-\epsilon(t-t_\mathrm{e})} - e^{-\epsilon(t_\mathrm{c}-t_\mathrm{e})} \right), & t_\mathrm{e} < t < t_\mathrm{c}, \\ 0, & t_\mathrm{c} < t < T_0, \end{cases} \tag{3.5}$$

where $\omega_\mathrm{g} = \pi / t_\mathrm{p}$ and $t_\mathrm{c} = T_0 = 1/f_0$. Parameters α and ϵ can be calculated from Equation 3.5 by assuming $g(t_\mathrm{e}) = E_\mathrm{e}$ and the energy balance $\int_0^T g(t) = 0$. Thus, parameters t_p, t_e, t_a, and E_e uniquely define the model. Another way of defining the model is presented in Fant et al. (1994), which is also linked to the voice source parameter R_d, proposed in Fant et al. (1994); Fant (1995, 1997). R_d effectively describes the functioning of the LF model in normal covariation of the full set of LF parameters. In addition to the Klatt and LF models, several other glottal flow models have been proposed (see, e.g., Rosenberg, 1971; Rothenberg et al., 1975; Hedelin, 1984; Ananthapadmanabha, 1984; Fujisaki and Ljungqvist, 1986).

Another way of estimating the parameter R_d based on phase minimisation, was proposed by Degottex et al. (2011) and further refined by Huber et al. (2012). The methods are based on estimating the parameter R_d of the LF model by minimising a phase-based criterion. The advantage of the method is that it does not require GIF or specific GCI-synchronous model fitting.

3.2.4 Frequency-domain parameters

The time-domain changes in the glottal flow, for example the changes in the phonation type from breathy to pressed, are also reflected as changes in the frequency domain. Therefore, several frequency-domain parameterization methods have also been proposed, especially those focusing on the quantification of the spectral decay of the glottal flow. This type of approach is beneficial in the sense that the processing can be made over several pitch periods, if desired, and thus no epoch detection is required. The source spectrum is typically computed using the fast Fourier transform (FFT), but also all-pole modelling has been used.

The simplest parameterization method for spectral skewness is the alpha ratio (Frøkjær-Jensen and Prytz, 1973), depicting the ratio between the spectral energies below and above a specific point in frequency. However, a more specific way of calculating the spectral decay of the glottal source is to utilise the amplitudes of the source harmonics. Childers and Lee (1991) have proposed a measure called the harmonic richness factor (HRF) to quantify the spectral decay of speech. The HRF is defined as the ratio between the sum of the amplitudes of the harmonics above the fundamental frequency (f_0) and the amplitude of the f_0 peak, that is

$$\mathrm{HRF} = \frac{\sum_{i \geq 2} H_i}{H_1}, \tag{3.6}$$

where H_i is the amplitude of the ith harmonic and H_1 is the amplitude of the f_0 peak. A similar but slightly simpler quotient, denoted by H1–H2, was proposed by Titze and Sundberg (1992) to measure the amplitude difference between the first and the second harmonics. H1–H2 has been widely used as a measure of voice quality. Also linear regression (Howell and Williams, 1988, 1992) and the parabolic spectral parameter (PSP) (Alku et al., 1997) have been proposed to model the spectral slope of the glottal flow. Alternatively, the glottal flow spectrum can be parameterized and modelled using a simple all-pole model (Raitio et al., 2011c). In addition to these glottal-flow-based features, there are methods that measure the contribution of the glottal source directly from the speech signal without using glottal inverse filtering (Hanson, 1997; Hanson and Chuang, 1999; Iseli et al., 2007). These methods, among other frequency-domain-based measures, are discussed in more detail in the study by Kreiman et al. (2007), in which the relationship between different parameters that measure the spectral decay are also studied. Finally, some methods directly create a mapping from frequency-domain parameters into glottal

features. For example, an artificial neural network is used to create a mapping from speech cepstrum to the OQ in Kane et al. (2013a).

As mentioned in Section 2.1.1, the glottal excitation also has an aperiodic component in addition to the quasi-periodic component. Characterising and quantifying this aperiodic component is useful in several applications, such as in speech modification and synthesis, where the degree of voicing and voice quality can be controlled by adjusting the amount of the aperiodic component. The simplest method for estimating the aperiodic component is to define a boundary frequency, often called the maximum voice frequency (F_{m}), which divides the spectrum into two parts, where the lower spectral band is dominated by the periodic excitation and the upper band is dominated by the aperiodic excitation. This idea is used in various vocoders (Griffin and Lim, 1988; Stylianou, 2001; Pantazis et al., 2008; Erro et al., 2014; Drugman et al., 2009b; Drugman and Dutoit, 2012) due to its simplicity and robustness. F_{m} can either be fixed (usually 4 kHz) for a given speaker and voice quality, as in the studies by Stylianou (2001), Drugman et al. (2009b), and Drugman and Dutoit (2012), or F_{m} can have a varying value from one frame to another, as in the study by Erro et al. (2014). Dynamic modelling of F_{m} has been observed to improve the naturalness of the synthesis in Drugman and Raitio (2014), which is Publication VI of this thesis. The estimation of F_{m} can be performed in several ways, such as using the peak-to-valley (P2V) measure calculated for all possible harmonic candidates (Stylianou, 2001), using a sinusoidal likeness measure (SLM) of the harmonic peaks (Erro et al., 2014), or by exploiting both the amplitude and phase information of the harmonics (Drugman and Stylianou, 2014).

Another way of quantifying the aperiodic part of the glottal flow is to use a multiband approach, where the energy ratio between the periodic and aperiodic components is estimated for each spectral band. These aperiodicity measurements can be computed in various ways. Yoshimura et al. (2001) calculate correlation coefficients for several frequency bands that define the amount of aperiodicity in each band. In the work by Kawahara et al. (2001) and Raitio et al. (2011a), aperiodicity, or the harmonic-to-noise ratio (HNR), is determined based on the ratio between the upper and lower smoothed spectral envelopes, which is defined by the amplitudes of the harmonic peaks and the interharmonic valleys, respectively. The ratios are then averaged across frequency bands in accordance with the equivalent rectangular bandwidth (ERB) scale (Moore and Glasberg,

1983) for perceptual weighting.

Usually the aim of the aperiodicity estimation is to quantify the amount of additive noise in the voice, which originates from the turbulent air at the glottis or above it. However, the aperiodicity of the glottal excitation may stem from other effects as well, such as jitter, shimmer, waveshape change, or some unknown combination of these factors (Murphy et al., 2008). Thus, more developed methods, such as glottal harmonics-to-noise ratio (GHNR) (Murphy et al., 2008), may be helpful for estimating the amount of the additive noise component.

3.3 Summary

This section described voice source estimation and parameterization. First, the concept of GIF was introduced, after which various GIF methods were described. Also the applications of GIF were shortly discussed. Next, GCI detection and polarity detection were introduced, which are required for many glottal flow parameterization algorithms. The parameterization of the estimated glottal flow was illustrated, categorising the parameterization methods into time-domain methods, glottal flow models, and frequency-domain methods. Finally, the estimation of the aperiodic part of the glottal flow was described.

4. Speech synthesis

The ultimate goal of text-to-speech (TTS) synthesis is to read any text and convert it to intelligible and natural sounding speech with desired speaker, contextual, and other extralinguistic characteristics such as emotion and speaking style. Although converting text to speech has a long history (see Section 4.1), meeting all of these goals simultaneously is extremely difficult. The complexity and difficulty of TTS synthesis can be easily illustrated by observing the corresponding process for a human speaker.

For humans, reading a text and speaking it aloud seems very effortless, although this process is very complex, and many of the phenomena in the brain are both functionally and physiologically not understood. In a very simplified and high-level view, reading a given text and speaking it aloud begins with the perception of the letters and words using our visual system: the eyes and the visual cortex. In the Wernicke's area in the brain (among other areas), the neural information is further processed for understanding the sentence and its meaning. Understanding the meaning of the text is important to produce a fluent and appropriate prosody for the utterance instead of just reading isolated words. Finally, the corresponding articulatory information is formed using, for example, the Broca's area in the brain and sent to the motor cortex and further to the articulators in the speech production mechanism. The movements of the articulators create speech sounds, forming a natural sounding and intelligible utterance with the voice characteristics of the speaker and having a convenient prosody and speaking style for the context. Accurate speech production, however, requires a constant feedback loop through the hearing mechanism in order to adapt and compensate our articulatory movements.

Although TTS synthesis is an extensive engineering problem, which requires sophisticated and complex algorithms, the process of converting

text into speech using a machine is less complex than the corresponding human process due to many simplifications. Firstly, machines are not yet capable of extracting high-level meaning from text, so this step is generally omitted in TTS synthesis. Secondly, speech generated by the physiological speech production mechanism is governed by the laws of physics as the air-flow from the lungs is modulated in the larynx and the vocal tract. This complex process is difficult to model accurately, and approximations are required in order to make speech synthesis feasible. These approximations result in less natural prosody and the possible misinterpretation of the meaning by the listener. Also, the approximations made at the signal level result in the loss of some of the perceptually relevant acoustic cues in speech, which may compromise naturalness, intelligibility, and extralinguistic factors of speech.

Despite the difficulty of speech synthesis, and the fact that we have billions of human speakers around us, the automatic generation of speech by machines has numerous existing and potential future applications—as already mentioned in the introduction, speech synthesis can be used anywhere where speech is to be used. Moreover, present-day high-quality TTS systems are approaching many of the qualities of human speech, which makes them useful and acceptable for increasingly many purposes. For example, speech synthesis has been widely used in assistive technologies to help people with a wide range of disabilities. TTS synthesis can be used in screen readers for people with mild or severe visual impairments. Screen readers can be implemented, for example, in computers and televisions to enable the user to listen to the speech generated from the displayed text. TTS is often used also as a communication aid for people with difficulties in producing speech due to, say, vocal disorders or dyslexia. For example, a speech-generating device (SGD) can be used as a personal device carried by the user to communicate with people through synthetic speech. Probably the most famous user of an SGD is Stephen Hawking, an English theoretical physicist and cosmologist who is also known for his work in popular science. In addition, a synthetic voice can be personalised for the user by voice banking (Yamagishi et al., 2012). Speech synthesis is also widely used in various telecommunication services, such as in information-retrieval systems. TTS synthesis also finds applications in computer-assisted language learning (CALL) (Handley, 2009), in the creation of audio books, talking toys, and entertainment products such as games and animations. In connection with other speech technologies,

such as automatic speech recognition (ASR), speech synthesis is also increasingly used in human-computer interaction (HCI), such as in mobile phones, where a user can give commands or ask questions, based on which the phone can give answers and feedback through synthetic speech (for example, Google Now, Apple's Siri, and Microsoft's Cortana). Finally, speech synthesis provides a fundamental tool for research in the production and perception of speech and language.

A system that can produce speech indistinguishable from a human with all the required expressivity is yet a distant dream, but already the current applications show that speech synthesis is becoming a part of people's lives, and most probably the use of speech synthesis will only increase with time. In the next section, the history and development of speech synthesis is briefly described, which shows that a machine that can imitate human speech has long been in people's minds, but only rather recently it has become widely useful in practice.

4.1 History of speech synthesis

In this section, the history and development of speech synthesis is briefly reviewed. The purpose of this section is not to give a complete and detailed discussion on the topic, but to give a general perspective and examples on the development of different synthesis techniques. More complete reviews on the development of speech synthesis can be found, for example, in Klatt (1987), Linggard (1985), and Flanagan (1976, 1972a,b).

The first successful attempts to produce artificial speech were based on imitating the speech production organs. In 1791, Wolfgang von Kempelen published a book presenting his speaking machine, which consisted of bellows (lungs), a vibrating reed (glottis), and a rubber tube modeling the vocal tract. The machine could produce many speech sounds if operated correctly. Several similar approaches have been studied since then (and probably even before), and today physical modelling of the vocal organs has applications from research to education. In the early 1920s, the research on speech synthesis started moving from physical modelling to using electric circuitry. Stewart (2012) built the first formant synthesiser, being able to produce static vowels. The system consists of two resonant circuits to model the two first formants which are excited by a buzzer. The first machine that could produce continuous speech was the vocoder by Dudley (1939), developed at the Bell Telephone Laborato-

ries. The device decomposed speech into slowly varying parameters and was able to reconstruct an approximation of speech from them. In the 1950s, the source-filter theory, summarised by Fant (1960), was applied to speech synthesis. The first such dynamically controlled synthesisers were Walter Lawrence's PAT (Lawrence, 1953) and Gunnar Fant's OVE I (Fant, 1953). Both of these synthesisers were later refined (Anthony and Lawrence, 1962; Fant and Martony, 1962), and they could produce a good approximation of human speech when the parameters were fine-tuned.

Starting from the 1960s, speech synthesis techniques split into two paradigms, articulatory-based synthesis (e.g., Kelly and Lochbaum, 1962) and signal-based synthesis. In articulatory synthesis, the physiological speech production mechanism is modelled more or less in detail. In signal-based synthesis, the speech signal itself is modelled, with any convenient means, in order to reconstruct the signal so that it is perceptually close to the original one. Although articulatory synthesis has not provided as good results as signal-based synthesis, it has been widely used to study speech production and perception, especially in conjunction with the latest measurement technologies, such as magnetic resonance imaging (MRI) and electromagnetic articulography (EMA). Recently, articulatory information has also been used in conjunction with the latest signal-based methods (see, e.g., Ling et al., 2009; Black et al., 2012; Astrinaki et al., 2013). However, signal-based methods, such as formant synthesis and LP synthesis provide better naturalness for synthesis applications. At the same time, several improvements in the modelling of the voice source waveform were also introduced. Instead of using simple (filtered) impulse trains or a sawtooth signal to model the glottal flow, detailed glottal flow models were introduced, such as the Rosenberg's model (Rosenberg, 1971), Klatt's model (Klatt, 1980), and the Liljencrants–Fant model (Fant et al., 1985a), providing improved flexibility and naturalness. Holmes (1973) also proposed the use of inverse-filtered glottal-flow waveforms to improve the naturalness of synthetic speech.

In parallel with the development of speech processing techniques, the generation of speech from text instead of mere speech analysis and synthesis was gaining more interest. Development work was done both at the signal and linguistic levels, and in 1968, the first full TTS synthesis system was developed by Umeda et al. (1968). Commercial speech synthesis products were introduced in the late 1970s, such as Dennis Klatt's MITalk formant synthesiser (Allen et al., 1987), which was later developed into

the product DECTalk. At that time, the first portable speech synthesisers appeared, such as the Synte 2 for Finnish in 1977 (Karjalainen et al., 1980) followed by the more widely known Speak & Spell toy by Texas Instruments, an example of the first mass production devices for speech synthesis.

In the 1980s, the paradigm started to shift from light and rule-based expert systems to database (corpus)- based systems, as speech data could be recorded, stored, and processed more efficiently. This was enabled by the reduced price in memory and increased computing power, and motivated by the potential to achieve better quality using concatenative synthesis methods. In concatenative speech synthesis, speech units (that can be of different size) are extracted from a speech database and concatenated at the synthesis stage according to specific rules or models. High-quality concatenative speech synthesis appeared in the 1990s, which was made possible again by the increased computational capacity and new methods in signal and natural language processing (Black and Taylor, 1994; Black and Campbell, 1995; Black and Taylor, 1997b; Hunt and Black, 1996). Purely software based synthesisers also became feasible for the same reasons, such as ATR's CHATR (Black and Taylor, 1994; Campbell and Black, 1996; Hunt and Black, 1996) and University of Edinburgh's Festival speech synthesis system (Black and Taylor, 1997a; Black et al., 2001). Unit selection-based systems provide rather natural speech quality. However, the concatenation of units may introduce some distortion, and thus the method may occasionally produce very low quality synthesis.

In the 1990s, statistical parametric speech synthesis (SPSS) using hidden Markov models (HMMs) (Tokuda et al., 1995b, 1999, 2002b; Zen et al., 2009; Tokuda et al., 2013) was introduced and has since then been the most researched paradigm in speech synthesis. The development of HMM-based speech synthesis was facilitated by the extensive research in ASR using HMMs since the same tools, such as the widely used hidden Markov model toolkit (HTK) (Young et al., 2006) can be utilised in both techniques. In SPSS, speech is converted into parameters that are statistically modelled and generated for synthesis using, for example, HMMs. The most widely used platform for SPSS is the HMM-based speech synthesis systems (HTS, 2014) developed in Japan.

Around the turn of the millennium, hybrid systems were introduced, which aim at combining the flexibility and robustness of SPSS and high segmental quality of unit selection synthesis. In the 2010s, deep learning

has been a strong trend in speech synthesis, and much of the research is still in progress (see Section 5.6). Deep learning utilises artificial neural networks with multiple hidden layers for the acoustic modelling, aiming at a better performance compared to the conventional HMM and Gaussian distribution-based acoustic modelling. Today, unit selection synthesis, SPSS, deep learning, and their various combinations constitute the state-of-the-art in TTS synthesis both for commercial and research purposes.

4.2 General TTS architecture

Language is the ability to express thoughts and ideas using a set of signs, whether acoustic, visual, or haptic. Speech is the oldest and most widely used comprehensive means of communication between people. In speech, different acoustic signals produced by the vocal organs are used as a set of signs of a language to convey information. Written text is another way of conveying information using a set of discrete symbols. There are, however, several differences between spoken language and written text. Spoken language is rich in contextual information; the way speech is produced depends largely on the speaker, whom it is directed to, what the context of the conversation or speech is, what the background knowledge of the speaker and listener(s) is, and so on. Speech is directed to someone at a specific time instant while written text can be read by anyone at any time and in any context. Due to these reasons, written text is only a shallow representation of spoken language[1], and the conversion from text to speech is not a trivial problem as it requires predicting several aspects of spoken language, such as pronunciation, prosody, and speaking style.

A TTS system aims to convert a text string into an acoustic speech pressure wave. A high-level flow chart of a speech synthesiser is shown in Figure 4.1. In general, a TTS system is composed of two parts: a linguistic or natural language processing (NLP) *front-end* and a DSP *back-end*. The front-end first normalises the text input to a standard written text, that is, it converts numbers and abbreviations into their full written forms. Then, the text is converted into a narrow phonetic transcription, which describes how the text should be pronounced, and generates information

[1]The relationship between the spoken and written language is not without a debate (Moxley, 1990), although written language is clearly more influenced by the spoken language.

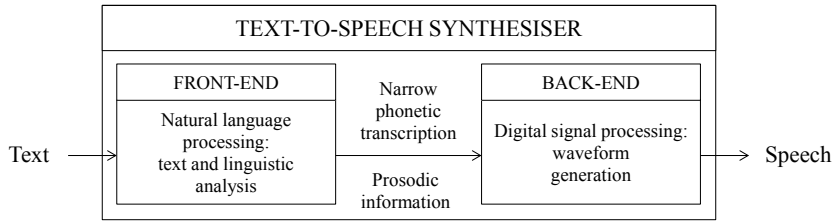


Figure 4.1. Functional diagram of a general TTS architecture.

for the intonation and rhythm, that is, the prosodic properties of the utterance. The back-end generates a speech waveform based on the specifications given by the front-end. In the following, the tasks and processes of the front-end and the back-end are described in more detail.

4.2.1 Front-end

Speech is composed of acoustic, phonetic, phonological, morphological, syntactic, semantic, and pragmatic levels of information (Dutoit, 1997). A written text is a shallow representation of speech, consisting of letters, numbers, symbols, punctuation marks, words, sentences, and paragraphs, indicating what and (approximately) how the text should be read by a human. For computers, however, such rules are not explicit enough, and the task of the front-end is to give exact specifications for the back-end on the characteristic of the speech signal to be generated.

The linguistic front-end first performs *text normalisation* or *pre-processing* to convert numbers, acronyms, and abbreviations into equivalent written-out words. This task is not straightforward since the correct conversion often depends on the context of appearance. After text normalisation, morphological and contextual analysis is applied, which categorises words into different word classes (also called part of speech (POS) or lexical class) in order to help predicting their pronunciation, which in many languages depends on the POS tag.

Based on previous information, the words are converted into a phonetic transcription which defines how the words are pronounced. This is also called letter-to-sound (LTS) conversion. In some languages, the LTS conversion is straightforward, such as in Finnish and Spanish, where each letter corresponds to a specific type of sound rather independently of the context[2]. In contrast, in many languages, such as in English and French,

[2]Despite possible simple pronunciation rules, the acoustic properties of the phonemes vary slightly depending on the context, which is handled in the latter stages of the speech synthesis process (see, e.g., Section 5.3.1).

the relationship between written and spoken languages is more complex, and the LTS conversion is more complicated due to the context-dependent nature of the LTS rules (Black et al., 1998), which usually also include a lot of exceptions. The LTS conversion can be performed either based on a pre-built lexicon, which consists of the list of words (or lexemes) of a given language and their pronunciations and POS tags, or using LTS rules for the conversion. Often both are used such that first the word pronunciation is sought from the lexicon, and if the word is not found[3], LTS rules are used to predict the pronunciation of the unknown word.

Finally, prosodic patterns of speech, such as phrasing and accentuation (manifested in the variation of pitch, intensity, segmental durations, and voice quality), are predicted by mapping linguistic and structural features of the text onto the patterns observed in speech. Each language has dedicated rules (although not explicit) on how the shallow textual information should be converted into the flow of speech. Simple heuristics or grammar-based systems can be used to parse the desired information for prosody generation. However, corpus-based modelling methods are more widely used in modern synthesisers, such as in unit selection and SPSS. For example, a prosodic model can be constructed by learning a mapping between the acoustic speech features of the corpus and the linguistic features (such as punctuation marks, POS tags, syntactic constituents, n-gram (Brown et al., 1992) distributions, etc.) extracted from the text. Thus, the learnt model or rules can be applied in the synthesis to predict phrases and accentuation from text. Alternatively, the whole statistical modelling framework of acoustic speech features (such as in SPSS) can be used to indirectly model the prosodic features based on the lower-level linguistic features.

Usually there is interaction between the difference phases of the front-end, since one phase cannot unambiguously guarantee a correct result, but outputs many possibilities of which the most probable is selected based on the information given by other phases. A more complete description of the linguistic front-end and various methods used for text parsing are described, for instance, in the work by Dutoit (1997) and Taylor (2009).

[3] It is practically impossible to construct a lexicon that contains all possible words and their pronunciations in a language. For example, the Google Web Trillion Word Corpus (Franz and Brants, 2006) contains 13,588,391 unique words, even after the words that appear less than 200 times are removed.

4.2.2 Back-end

The task of the back-end is to convert the symbolic linguistic representation given by the front-end into speech sounds. In many TTS systems, this part also includes the computation of the target prosody (pitch, intensity, duration, etc.), which is used for synthesis. The back-end may be based on several different methods to generate the actual speech waveform, such as concatenation of speech units, or to utilise statistical modelling and the source-filter based vocoder for generating speech from parameters. Although the back-end relies on the information given by the front-end, much of the naturalness and intelligibility depends on the type and implementation of the back-end. In the following section, different back-end synthesis methods are described in more detail.

4.3 Speech synthesis methods

The front-end is often rather language dependent while the back-end methodology is more independent of language—in corpus-based methods, the back-end is only trained with the speech material of the given language. Thus, a linguistic front-end for a language can be used for driving various types of back-end synthesis methods. If the LTS conversion in the front-end works well, the most important contribution of the front-end after that is to provide appropriate prosody, or appropriate contextual information for the back-end to generate the prosodic properties of speech. The modelling of prosody is considered hard, and the topic is partly outside the scope of this thesis. The segmental speech quality and various voice characteristics of synthesised speech, however, are highly dependent on the back-end synthesis method. Therefore, TTS systems are often classified based on the back-end type into different *synthesis methods*, which are described in the following sections.

4.3.1 Formant synthesis

Formant synthesis, also known as rule-based synthesis, is based on using a set of parameters to create speech, such as formant frequencies, amplitudes, and bandwidths as well as fundamental frequency, voicing, and amount of aspiration noise. The voice source of speech is often modelled as a series of pulses or noise, and the formants are modelled as individ-

ual formant filters connected in parallel, series, or both. Speech is produced by creating rules on how the parameters vary with respect to time based on the input from the front-end. The rules are often created manually by human experts, which often also requires laborious trial-and-error work by comparing the original and synthesised utterances to optimise the rules.

Formant synthesis is one of the oldest synthesis techniques, and it received much attention through Dennis Klatt's publication of a sophisticated formant synthesiser (Klatt, 1980, 1982). The naturalness of formant synthesis is generally rather poor due to the limited set of rules constructed by human experts and due to the overly simplified synthesis techniques. However, formant synthesis can be very intelligible even at very high speaking rates, as it avoids the acoustic glitches that commonly occur in concatenative systems. Theoretically, formant synthesis can produce high-quality synthetic speech, as was shown by Holmes (1973) by creating a synthetic speech sample sounding so natural that an average listener could not tell the difference between the synthetic and natural samples (Klatt, 1987). However, in practice, it is difficult to build a comprehensive set of rules in order to yield high-quality TTS synthesis. Despite the weaknesses, formant synthesis has applications in, for example, reading machines for the disabled and in speech research (Pickett, 1999).

4.3.2 Articulatory synthesis

Articulatory synthesis aims to model the speech production mechanism as accurately as practically possible. Theoretically, articulatory synthesis is able to produce very natural sounding speech if an accurate enough model is used. However, for practical reasons, such as limitations in the current speech production models and computational power, articulatory synthesis has not achieved much success compared to other speech synthesis methods. Articulatory synthesis is, however, useful in basic speech research, and articulatory features obtained through the latest measurement technologies, such as magnetic resonance imaging and (MRI) and electromagnetic articulography (EMA), have been used in modern statistical parametric speech synthesisers (Ling et al., 2009; Black et al., 2012; Astrinaki et al., 2013).

4.3.3 Linear prediction synthesis

Similar to formant synthesis, LP synthesis is based on time-varying speech parameters, but LP filter coefficients can be automatically estimated from a short-time frame of the speech signal instead of manually finding the parameters for individual formant filters. LP-based synthesis utilises the source-filter model of speech production (Fant, 1960), which was reviewed in Section 2.4. Thus LP synthesis is based on two components, a driving excitation signal and a time-varying filter, as is depicted in Equation 2.3. In LP analysis, the spectral contributions of the vocal tract filter and the glottal voice source are both captured by the estimated LP filter, and the excitation signal becomes white. Therefore, the driving excitation in LP synthesis in the simplest form consists of impulses at the glottal closure instant for voiced frames and white noise for unvoiced frames.

LP is a widely used method in speech technology (for a review, see, for example, Makhoul, 1975; Rabiner and Schafer, 1978; Markel and Gray, 1980), and its usefulness is based on its accuracy to estimate the spectral envelope of speech and in its relative speed of computation. LP coefficients also have good interpolation and smoothing properties when converted to, for example, the linear spectral pair (LSP) representation (Soong and Juang, 1984). The weaknesses of LP synthesis are its inability to model spectral valleys and the possible bias in the estimated speech spectrum due to the voice source harmonics. LP-based synthesis also suffers from buzziness if a simple excitation model consisting only of impulses and white noise is used. Today, LP-based synthesis techniques are not commonly used in rule-based synthesis but in conjunction with other speech synthesis techniques, such as for representing spectral information in statistical parametric speech synthesis or for speech compression and modification in concatenative synthesis.

4.3.4 Concatenative synthesis

Concatenative synthesis provides a different approach to speech synthesis. Instead of artificially generating speech, a prerecorded speech corpus is first split into (small) speech segments, which are then concatenated smoothly in the synthesis phase to generate new sentences. Generally, concatenative synthesis provides high-quality speech output, but it may often suffer from audible glitches in the output due to imperfect concate-

nation of the units.

There are three subtypes of concatenative speech synthesis that use different types of units in concatenation. In *domain-specific synthesis*, recorded words and phrases are used to create complete utterances. This type of approach can be used in applications were the desired speech output is limited to a small specific domain so that all possible outputs can be generated with a reasonable effort. Such applications are, for example, scheduled announcements in public transportation and weather reports. The synthesis quality can be very high due to the natural speech recordings, but often the prosody of the concatenated utterances can be rather poor.

In *diphone synthesis* (see, e.g., Dutoit et al., 1996), a minimal speech database is constructed that contains all the diphones (phone-to-phone transitions) occurring in a language. In diphone synthesis, only one example of each diphone is contained in the speech database, and the number of diphones depends on the language. In synthesis, the diphones are concatenated with the aid of signal processing methods, such as LP and pitch-synchronous overlap-add (PSOLA) (Charpentier and Stella, 1986; Moulines and Charpentier, 1990). However, diphone synthesis often suffers from glitches when concatenating two diphones that are not compatible with each other. The naturalness of diphone synthesis can also suffer from artefacts stemming from signal processing methods, which are applied in order to compress the diphone inventory or modify the diphones in order to aid the concatenation.

Unit selection synthesis is today the most common concatenative speech synthesis method, and probably the most widely used commercial synthesis method of all. Unit selection systems are usually based on a large speech corpus, which is segmented into units that can be of various length: frames, half-phones, phones, diphones, triphones, demisyllables, syllables, morphemes, words, phrases, sentences, or a combination of these (Breen and Jackson, 1998; Segi et al., 2004). Most commonly the units are rather small (not words, phrases, or sentences) in order to preserve the flexibility of the synthesis. The speech database is segmented into the units using a specially designed speech recogniser that performs *force alignment*, and possibly checked and corrected manually. The speech units are then indexed, clustered, and labelled according to linguistic and acoustic features, such as phone identity, f_0, energy, spectrum, and duration, and contextual information, such as position in the syllable, word,

and phrase, and neighbouring phones (Black and Taylor, 1997b). In the synthesis, optimal target units for a sentence are selected by minimising target and concatenation costs (Black and Campbell, 1995; Hunt and Black, 1996; Black and Taylor, 1997b). The target cost defines how well a unit matches the linguistic and acoustic features and context provided by the front-end. The concatenation cost measures how well adjacent units can be concatenated and is evaluated by comparing the acoustic features of the two units at the concatenation point. The weights of the target and concatenation costs and different features thereof can be either hand-tuned or more often automatically trained to optimise the synthesis quality (Black and Campbell, 1995; Hunt and Black, 1996). Also, signal processing techniques called smoothing algorithms can be used to aid the concatenation if a suitable unit is not found.

Unit selection synthesis can provide highly natural and intelligible speech if a large (single-speaker) corpus is used and the system is well optimised. Best unit selection systems can be indistinguishable from human speech in the style the system is optimised for. However, the concatenation points may still cause audible glitches or anomalies in prosody, especially when using smaller corpora. Despite the high quality, only a single speaking style, usually read-aloud, can be produced using one extensive speech corpus. Sampling-based approaches, such as unit selection, are inherently inflexible and limited by the available samples in the database, which limits the ability of the system to change, for example, voice quality, speaking style, or expression (Black, 2003). Designing and recording a unit selection corpus that includes all the desired variation is highly impractical. Unit selection synthesis also requires large data storage for the recorded units, which may limit its use in some applications, although this limitation is likely to be less important with decreasing prices and increasing capacity of memory and data storage.

4.3.5 Statistical parametric speech synthesis

SPSS (Zen et al., 2009; Tokuda et al., 2013) uses principle similar to formant synthesis, that is, parameterizing and reconstructing speech with whatever means works best, but the parameters are estimated automatically from a speech corpus and modelled statistically for representing speech sounds in different contexts. Usually, a source-filter model is utilised to represent speech signal as a set of excitation and spectral features, although other parameterization methods can be also used, such as

the harmonic plus noise model (Erro et al., 2014; Degottex and Erro, 2014) or the dynamic sinusoidal model (Hu et al., 2014). Usually decision-tree clustered context-dependent HMMs are utilised for modelling the time-varying speech parameters, and thus SPSS is often called *HMM-based speech synthesis*. However, the latter is only one instance of SPSS, since also, for instance, deep neural network (DNN) can be used instead.

SPSS consists of two phases. First, in the analysis stage, a dedicated vocoder is used to decompose the speech corpus into speech parameters. In HMM-based speech synthesis, each context-dependent phoneme (defined in the front-end) is represented with a left-to-right HMM, which represents the properties of the phoneme by modelling each acoustic parameter with a Gaussian distribution in each state. The states of the context-dependent phonemes are clustered using a decision tree (discussed in Section 5.3.1) in order to handle data sparsity. In the synthesis stage, a phoneme sequence given by the front-end is created by concatenating context-dependent HMMs from which smooth parameter trajectories can be generated using the means and variances of the Gaussians. Finally, the vocoder is used to generate the speech waveform from the parameter trajectories. The detailed methodology of HMM-based speech synthesis is discussed in Section 5.

Since SPSS uses a parametric form of speech, it can generate smooth and intelligible speech. Unlike the unit selection method, SPSS is able to generate speech that is not included in the original corpus by predicting the parameter values for a new context. SPSS is also flexible in the sense that it can be adapted (Yamagishi et al., 2009a) to a different voice quality, speaking style, or speaker identity by using a small amount of corresponding speech material. SPSS does not require as large a speech database as the unit selection methods, and the footprint of SPSS is very small; the statistical model required only a fraction of the size of a large corpus. However, due to the parametric representation of speech, SPSS suffers from lower segmental speech quality than unit selection synthesis. The generated parameter trajectories and the spectrum are also over-smooth due to averaging in the statistical modelling, which degrades the synthesis quality. Recently, several improvements have been introduced in SPSS to improve the synthesis quality, which makes SPSS acceptable even in commercial use.

4.3.6 Hybrid methods

Although unit selection and statistical parametric speech synthesis use completely different waveform generation, they also have a lot in common. Clustering methods used in unit selection synthesis are very similar to the ones in HMM-based synthesis, the main difference being the representation of clustering: the statistics of the context-dependent HMMs or the multi-templates of speech segments. Also, the decision trees used in SPSS (discussed in Section 5.3.1) are essentially equivalent to the regression trees in unit selection systems. The likelihoods of static and dynamic features in SPSS also correspond to the target and concatenation costs in unit selection synthesis.

Due to these similarities, there are several hybrid approaches where the benefits of both methods are exploited (Zen et al., 2009). For example, the HMM-based approach has been used in unit selection synthesis to predict targets units or calculating costs (Kawai et al., 2004; Hirai and Tenpaku, 2004; Rouibia and Rosec, 2005; Yang et al., 2006; Krstulović et al., 2007; Lu et al., 2009; Qian et al., 2010; Jiang et al., 2010; Chen et al., 2011, 2013; Zhang et al., 2009; Meen and Svendsen, 2010; Yu et al., 2007; Ling et al., 2008; Yu et al., 2013b; Huang et al., 1996; Hon et al., 1998; Kominek and Black, 2006; Okubo et al., 2006; Ling and Wang, 2006, 2007; Ling et al., 2007), and guiding smoothing (Plumpe et al., 1998; Wouters and Macon, 2000). Mixing natural and synthetic units in multi-form speech synthesis has also been proposed in order to fix data sparsity (Okubo et al., 2006; Aylett and Yamagishi, 2008; Pollet and Breen, 2008; Sorin et al., 2011; Tiomkin et al., 2011; Sorin et al., 2014). Also, the complete unification of both approaches has been investigated (Taylor, 2006).

Hybrid approaches provide several benefits compared to either plain unit selection or HMM-based synthesis. The over-smoothing problem of SPSS and quality degradation due to vocoding is avoided by using natural units. The HMM-based costs also help to capture the detailed context-dependencies. On the other hand, hybrid approaches lose the flexibility and small footprint by using the natural units.

4.4 Evaluation of synthetic speech

A great thing in speech synthesis research is that the results can be assessed by listening to the synthesis output. This is commonly done

by speech researchers when developing new methods for TTS synthesis. However, in order to guarantee consistent comparison and improvement of TTS systems, the synthesis output must be evaluated in a more formal manner. The evaluation of synthetic speech is not a trivial problem for several reasons. Firstly, there is no absolute reference speech waveform to compare with, since the TTS process means generating new instances of speech. Even if the same utterance is spoken by a human speaker and generated by a TTS system, the two speech waveforms would be different with respect to many aspects, such as prosody and duration, and thus objective measures, such as used in speech coding, cannot be generally used. Secondly, there are multiple criteria on how to assess synthetic speech (see, e.g., Mayo et al., 2005, 2011; Hinterleitner et al., 2011, 2013). A TTS system aims to produce both natural and intelligible synthetic speech that should also represent speaker characteristics, expressions, and contextual cues. Various TTS techniques may have a different degree of success in achieving each of these goals, and depending on the desired application, emphasis on different aspects may be used. Thirdly, usually the best means for assessing synthetic speech is subjective testing, which is usually time-consuming and expensive. In order to yield useful results, evaluation criteria, question setting, and the evaluation methodology must be carefully considered. Subjective assessment also requires special arrangements, such as appropriate equipment for the test and native speakers of the language, who are also influenced by their familiarity with the evaluation methodology, the possibility for rehearsal, and the listening conditions. Finally, a statistical analysis of the evaluation results must be performed in order to draw meaningful conclusions. Through careful and rigorous test design and assessment, useful and reliable results can be achieved for developing new techniques both in the TTS front-end and back-end.

Since a speech synthesis system consists of various components which themselves can be highly complex, an engineering point of view is often adopted for evaluating synthetic speech. Although single TTS system components can be assessed individually (for example, text normalisation in the front-end), usually a TTS system can be seen as a black box that outputs speech, which is the final measurable quantity. In general, synthetic speech is assessed by its naturalness, intelligibility, or extralinguistic characteristics, such as speaker similarity or identity, expressivity, and suitability for the context. In the following, the evaluation of these

Table 4.1. Rating scale for overall speech quality (International Telecommunication Union, 1996).

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

criteria is discussed further.

4.4.1 Evaluation of naturalness

Naturalness[4] of synthetic speech is a very multi-dimensional concept since various aspects in synthetic speech can make it sound unnatural to a human listener; both the linguistic front-end and the signal processing back-end may produce their own artefacts, such as mispronunciation, unexpected prosody, or signal artefacts and robotic sound quality. Nevertheless, humans have a good idea of how natural speech sounds, and therefore naturalness of synthetic speech is often assessed by humans in subjective listening tests.

Naturalness can be assessed, for example, by playing synthetic speech samples to listeners and asking them to rate the samples on a Likert scale (Likert, 1932) from 1 to 5, where each number is given a verbal description. In the evaluation of overall naturalness (or quality, as defined in International Telecommunication Union, 1996), the scale in Table 4.1 is commonly used. This type of test is also known as the absolute category rating (ACR) test (International Telecommunication Union, 1996). Averaging the scores results in the mean opinion score (MOS), which indicates the quality of each evaluated system. This usually gives a reasonably accurate figure for systems' performance if enough listeners are used in the evaluation. For example, such a test is used in Publication IX. However, small differences between systems is harder to assess using the MOS. Also, the interpretation of the MOS score is harder—it is difficult to determine which aspects listeners paid attention to when assessing the

[4]Speech *naturalness* and speech *quality* are often considered to be comparable terms, and both are used to describe the overall impression on synthetic speech. In this thesis, both terms are used depending on the context, but it is important not to confuse them with the term *voice quality*, which is discussed in Section 2.3. For a discussion on the perceptual quality dimensions of TTS systems, see, e.g., Mayo et al. (2005, 2011) and Hinterleitner et al. (2011, 2013).

Table 4.2. Preference rating scale between two systems when comparing the quality of the second utterance to the quality of the first (International Telecommunication Union, 1996).

Quality of the second sample compared to that of the first	Score
Much better	3
Better	2
Slightly better	1
About the same	0
Slightly worse	−1
Worse	−2
Much worse	−3

samples. MOS tests do not necessarily give absolute results that are comparable between evaluations performed at different times and in different conditions. Especially the systems included in a test greatly affect the results. For example, if natural speech is included in a test, the synthetic speech samples will be rated worse than without a natural reference.

In order to compare two or more systems directly and to assess smaller differences between systems, a comparison test can be used. In such a test, test subjects listen to two samples, one from each system, and choose the sample they prefer or rate the quality difference between the two samples. If the test subject is instructed to select the preferred sample, a preference score is obtained by calculating the percentage of how often a system was preferred over the other. Such preference tests are used in Publications IV and IX. If the objective is to measure the quality difference between two or more systems, a comparison category rating (CCR) test (International Telecommunication Union, 1996) methodology can be used, which utilises a discrete seven-point scale ranging from −3 (much worse) to 3 (much better), which is presented in Table 4.2. The results are averaged to obtain the comparison mean opinion score (CMOS) for each system, which gives the order of preference of each system and the amount by which the quality differs. Such tests are used in Publications III, IV, V, VI, and VIII.

As stated earlier, naturalness is a complex concept, and rating naturalness depends on several other things than might be obvious at first: factors such as expectations and experience of the test subject and the context of the application may have significant effects on the final impression. For example, the final “naturalness” can be assessed only in the end product where the synthesiser is used, whether it is a human-computer interaction using a mobile phone or an animated character in a movie. In

such situations, likability or believability (Campbell, 2007) may be better descriptors for the goodness of a system to the user.

Finally, verbal, written, or any other type of feedback from listeners or real users may be crucial for developing a TTS system (see, e.g., Lu et al., 2011). Especially if the user feedback and corrections can be performed in an automated fashion, the TTS system can be constantly improved (see, e.g., Simple4All, 2014).

4.4.2 Evaluation of intelligibility

Speech intelligibility can be measured either by evaluating overall speech comprehension or by evaluating the recognition of single speech segments, phonemes or words in isolation or in a sentence. Comprehension tests evaluate how well a message delivered by a synthetic utterance is understood by the listeners (or alternatively how much time or resources speech comprehension takes). Comprehension of speech might be one of the key aspects in developing a TTS synthesiser—after all, the final goal of a TTS system is to deliver a message to a listeners as accurately and with as little effort as possible. Some studies have shown that the comprehension of synthetic speech is more difficult compared to natural speech (Duffy and Pisoni, 1992). On the other hand, the results by Paris et al. (1995) and Chang (2011) show that the comprehension of synthetic speech is not worse than that of natural speech. The discrepancy between the results may be explained by the evaluation methodology. In general, conventional comprehension tests may not be appropriate for assessing real life speech comprehension, where *attention* is a significant factor, influenced by the naturalness, prosody, and possible artefacts of synthetic speech. Thus laboratory-based comprehension tests, where attention is maximised by definition, may have to be replaced with more realistic comprehension test scenarios. Nevertheless, intelligibility tests based on, say, word recognition are easier to conduct than speech comprehension tests, and these methods, discussed below, have shown their power in the systematic assessment of speech synthesisers and their components.

The earliest intelligibility evaluation methods based on single segments, phonemes, or phoneme combinations are rarely used today for evaluating complete TTS systems. Instead, word recognition tests are commonly used for evaluating the intelligibility of a TTS system. In a word recognition test, words are played either in isolation or in sentences to subjects who are requested to indicate what they have heard. The intelligibility is

measured by the word error rate (WER) evaluated from the answers. For example, in the (diagnostic) rhyme test (DRT) (Fairbanks, 1958; Voiers, 1977) and the modified rhyme test (MRT) (House et al., 1965), similar sounding words are played to the listeners, such as BAD, BACK, BAN, BASS, BAT, BATH. Intelligibility testing at the sentence level can be performed using for example Harvard sentences (IEEE Subcommittee on Subjective Measurements, 1969), such as "JAZZ AND SWING FANS LIKE FAST MUSIC", which are designed to represent the natural distribution of phonemes in English. Such a test in Finnish is used in Publication VII.

However, words in semantically sound sentences are rather easy to guess even if a word is not properly heard. Depending on the selected sentences, this may result in a very low WER with good synthesisers, which limits the discrimination ability of the test. Therefore, semantically unpredictable sentences (SUS) are often used in order to prevent guessing. SUS sentences are grammatically correct so that they form valid sentences, but the sentences do not necessarily make sense, for example, "THE PLANE CLOSED THE FISH THAT LIVED". A SUS test (Benoît et al., 1996) can be used to obtain more reliable intelligibility scores. Such test in Finnish is used in Publication III. Alternatively, the flooring (or ceiling) effect of such tests can be avoided by using a speech reception threshold (SRT) test (Plomp and Mimpen, 1979; Vainio et al., 2005). In the SRT test, the test material is presented in noise with an adaptive SNR, and the intelligibility is defined by the SNR at which a specific WER, usually 50%, is obtained.

Often intelligibility in a specific environment can be of interest as well. Various noise types affect the intelligibility of speech, and even the spatial noise distribution has some effect on speech intelligibility (Raitio et al., 2012a). By modifying the speech output, for instance, by simulating Lombard speech (Raitio et al., 2011b, 2014c), improved intelligibility in noise can be achieved. An intelligibility test in noise with natural sentences is used in Publication VII.

4.4.3 Evaluation of extralinguistic characteristics

In addition to naturalness and intelligibility, speech carries a lot of extralinguistic information, such as speaker identity, expressivity, and various other voice characteristics. Often speaker similarity is assessed in order to test how well a synthesiser can reproduce specific speaker traits. Speaker similarity can be evaluated using an approach similar to that

used in the evaluation of naturalness by changing the question setting and descriptions in the verbal scales. For example, synthetic speech samples can be presented to subjects with a (different) natural reference sample, and they are asked how similar or different the speaker in the test sample sounds compared to the reference sample using a Likert scale ranging from 1 (totally different person) to 5 (exactly the same person). Alternatively, if two or more methods are compared in terms of similarity, an ABX test can be performed so that the subject is presented with two samples and instructed to select the one (A or B) that is more similar to the reference natural speech (X). Such an approach is used in Publications IV and V. A similar approach can be used to measure virtually any characteristic of speech, such as suitability and likability of different speaking styles in different sound environments (Publication VII), impression of shouting and use of vocal effort (Publication VIII), and impression of creakiness (Publication IX).

The evaluation of prosody is especially difficult since it builds on several characteristics of speech (Sonntag and Portele, 1998). While evaluating one dimension of speech, other dimensions may easily affect the results, especially when the assessment extends from the evaluation of individual speech segments to the analysis of full paragraphs. Even if listeners are asked to rate specific prosodic features, they may have difficulties in pinpointing specific causes for their perception. For example, perceptual rating of intonation is affected by the segmental quality of speech (Vainio et al., 2002). Moreover, there is no such thing as "correct" prosody, which complicates the evaluation even more.

There are no established standards in prosody evaluation, and thus general assessments have been conducted, such as measuring the acceptance of a human-machine dialogue by simulating the dialogue situation (Pols, 1989), which, however, is unable to provide diagnostic information. A similar approach was adopted in the Blizzard Challenge 2012 for evaluating the prosody of audio-book reading (King and Karaiskos, 2012). Other experiments on the evaluation of prosody have utilised, for example, delexicalisation (Sonntag and Portele, 1998; Vainio et al., 2009), objective assessment (Hirst et al., 1998), and eye-tracking (Rajkumar et al., 2010) in addition to conventional methods.

4.4.4 Objective evaluation

Since subjective experiments are expensive, time consuming, and often not reproducible, automatic objective evaluation would offer an attractive alternative for assessing synthetic speech. However, objective evaluation of synthetic speech is often not possible due to the lack of a reference signal—whereas in speech enhancement and coding, the original unprocessed speech signal can be used as a reference, synthetic speech is created from text and thus there is no reference speech waveform available.

However, in certain cases, a reference natural speech waveform can be used. For example, synthetic speech can be generated according to the specifications based on natural speech samples. If the phonetic and contextual labelling and duration information is first extracted from natural speech and then used to generate synthetic speech, the original natural speech sample can be used as an approximate reference signal for the synthetic sample. Various methods can then be used to compare the synthetic signal to the natural reference. A possible problem in this sort of evaluation is that even slight differences in alignment and duration may result in quite different objective scores.

Common objective measures include various distortion measures based on the speech spectrum, such as log-spectral distortion (LSD) (Gray and Markel, 1976; Gray et al., 1980), cepstral distance measure (Gray and Markel, 1976; Gray et al., 1980), mel-cepstral distortion (Kubichek, 1993; Toda et al., 2007), Itakura–Saito distortion (Itakura and Saito, 1968; Gray and Markel, 1976; Gray et al., 1980), and Kullback–Leibler distance (Veldhuis and Klabbers, 2003), which all give a measure of how different the synthesised speech signal is from the original speech signal in terms of the spectrum. Similarly, measures such as root mean squared error (RMSE) (Clark and Dusterhoff, 1999), correlation (Clark and Dusterhoff, 1999), and log likelihood ratio (Lu et al., 2010) can be used to compare the speech parameters generated by a synthesiser to the ones extracted from the original speech file, or alternatively their probability distributions using, for example, Kullback–Leibler divergence (Do et al., 2014). Commonly used speech features in evaluations are, for example, $\log f_0$, intensity, and various spectrum-based parameters. Also the voiced/unvoiced error rate is often used as a measure of quality. The SNR in comparison to original speech file can also be used to indicate quality differences, such as is used in segmental SNR and frequency-weighted segmental SNR (Hu and

Loizou, 2008). However, the problem with all of these methods is that they are often less correlated with human perception, and thus they can be used only as an indication of perceptual difference or improvement.

Objective methods that aim to model human perception have also been developed, such as perceptual evaluation of speech quality (PESQ) (International Telecommunication Union, 1997; Rix et al., 2000, 2001), which was originally developed for evaluating speech quality in speech transmission technology. PESQ estimates the MOS for a given speech signal when compared to a natural reference. PESQ first converts both speech signals into features that correlate with perception and then maps the parametric difference to a MOS scale.

Similar methods can be used to objectively estimate the intelligibility of synthetic speech, although this task is even harder than estimating the naturalness. Usually objective intelligibility methods measure the intelligibility in the presence of noise. The most commonly used measures are the articulation index (AI) (French and Steinberg, 1947; Fletcher and Galt, 1950; Kryter, 1962), speech transmission index (STI) (Steeneken and Houtgast, 1980), speech intelligibility index (SII) (American National Standards Institute, 1997) or its modification, coherence SII (Kates and Arehart, 2004), Dau measure (Christiansen et al., 2010; Dau et al., 1996), glimpse proportion (GP) (Cooke, 2006), and short-time objective intelligibility measure (STOI) (Taal et al., 2011). PESQ has also been used to estimate the intelligibility of speech (Beerends et al., 2004, 2005). Taal et al. (2009) and Hu and Loizou (2008) have evaluated various measures for speech intelligibility prediction in noise. Valentini-Botinhao et al. (2011a,b) have measured the correlation of multiple objective measures with respect to subjective intelligibility ratings with synthetic speech, concluding that the Dau measure and glimpse proportion had the highest correlation. The correlating of subjective results and PESQ with synthetic speech has been studied with promising preliminary results by Cerňak and Rusko (2005). Also, the relation of subjective and objective scores for prosody evaluation has been studied by Hirst et al. (1998).

Due to the correlation of subjective and objective evaluation methods, attempts have been made to construct an evaluation method based on large subjective evaluation data, for example, from the Blizzard Challenge (2014). Such objective measures have shown to provide promising results (Falk et al., 2008; Hinterleitner et al., 2010; Huang, 2011). Also speech recognition has been used to assess synthetic speech without the need for

a natural reference signal (Cerňak et al., 2009).

4.4.5 Public speech synthesis evaluations

Finally, in order to achieve comparable evaluation results for various synthesis techniques, the testing conditions should be ideally the same. Differences in the speech database, evaluation method, listening test setup, listening conditions, and listeners may lead to different results. Also, if only the signal processing back-end is evaluated, the same front-end should be used as the results may highly depend on it. This has long been a problem in evaluating speech synthesis, and it still exists. Using generally known and publicly available reference methods for comparison is widely used, but still differences due to settings, corpora, and testing conditions may lead to different results. Only recently, a public speech synthesis assessment has been devised called the Blizzard Challenge (Black and Tokuda, 2005; Blizzard Challenge, 2014; King, 2014). Blizzard Challenge is an annual open assessment where speech synthesis entries are built and evaluated using the same data and a large subjective evaluation. Blizzard Challenge has been organised since 2005, and it has been recognised as a reliable and valuable benchmark for speech synthesis techniques. Various speech materials for building the voices have been used from large to small corpora in many different languages and ranging from the read-aloud TTS style to audio-book data. Various evaluation methods and criteria have also been used. For example, speech naturalness, intelligibility, speaker similarity, and intelligibility in the presence of noise have been assessed. Blizzard Challenge has raised a lot of awareness on speech synthesis and its evaluation, and it has given valuable information on speech synthesis techniques and evaluation, which are summarised by King (2014). Blizzard Challenge has also created spin-offs, such as the Albaysin Challenge (Campillo et al., 2011), which is similar to the Blizzard Challenge but uses Spanish speech data, and the Hurricane Challenge (Cooke et al., 2013), which evaluates the intelligibility of natural and synthetic speech in noisy conditions. The vocoder and related techniques presented in this thesis have been used in several submissions for the Blizzard Challenge evaluation (Suni et al., 2010, 2011, 2012; Watts et al., 2013; Suni et al., 2014). For example, the system described in Suni et al. (2010) was the most intelligible in the presence of noise among all systems, even more intelligible than natural speech. Also, the system described in Suni et al. (2014), using DNN-based voice source modelling

(Raitio et al., 2014a,b), was very successful among parametric speech synthesisers with the six different Indian languages.

4.5 Summary

This section described the fundamental concepts in speech synthesis and briefly reviewed the history and development of speech synthesis techniques. The general architecture of a TTS system and the roles of the linguistic front-end and the signal processing back-end were described. Various techniques for speech synthesis were presented with their advantages and disadvantages. Finally, the problem of evaluating synthetic speech was discussed, and various methodologies for speech synthesis evaluation, both subjective and objective, were presented, concentrating on different aspects of synthetic speech, namely naturalness, intelligibility, and extralinguistic characteristics.

5. Statistical parametric speech synthesis

SPSS (Black et al., 2007; Zen et al., 2009; Tokuda et al., 2013), already briefly introduced in Section 4.3.5, is one of the most widely used speech synthesis technologies today. Although SPSS does not yet provide as good naturalness as the best unit selection methods, its flexibility (Yamagishi et al., 2009a) and robustness (Yamagishi et al., 2009b) makes it an attractive method for almost any speech synthesis application. Moreover, the speech quality of SPSS has improved a lot during the last decade, and currently the quality of SPSS has reached a level where it can stand in its own right (Zen et al., 2009).

The idea of SPSS is to utilise a parametric representation of speech, using a vocoder that can convert speech into a meaningful set of parameters that describe the perceptually most important characteristics of speech. Using the parametric representation of speech and the linguistic information extracted from text, the context-dependent statistics of the speech sounds can be modelled. Usually decision-tree clustered context-dependent HMMs are utilised for the statistical modelling of the time-varying speech parameters, and thus SPSS is often called *HMM-based speech synthesis*. In synthesis, new speech parameters can be generated according to the text input, and the parameters can be fed back to the vocoder to reconstruct the speech signal.

This simple principle has led to a totally new paradigm in speech synthesis, which is more thoroughly described in this section. First, the fundamental principles and architecture of HMM-based speech synthesis are reviewed, after which SPSS is discussed in more detail, concentrating on topics such as flexibility and adaptation, various vocoders, and voice source modelling. Finally, possible future directions of SPSS are discussed.

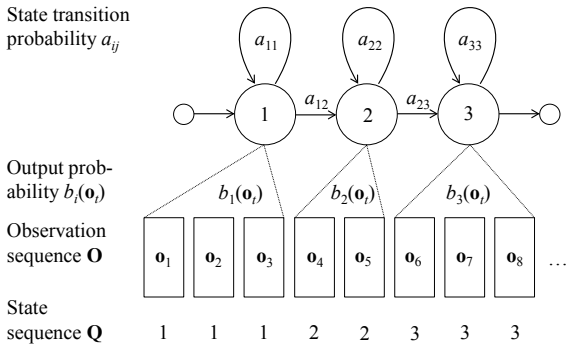


Figure 5.1. Example of a three-state left-to-right HMM. The states of the HMM are denoted with large circles numbered from one to three. A state transition probability from state i to state j is denoted as a_{ij}. An output probability density of state i is denoted as b_i, and the generated observation at time instant t is denoted as $\mathbf{o}_t$. Adapted from Tokuda et al. (2013).

5.1 Hidden Markov model

The HMM is a powerful statistical tool for the modelling, segmentation, and classification of a discrete time-series. The basic theory of the HMM was first published in a series of papers by Baum and Eagon (1967) and Baum et al. (1970), and today HMMs are widely used in various applications. HMMs have been successfully used in speech and language processing, such as in speech recognition, enhancement, and synthesis as well as in language modelling, translation, and understanding.

A Markov chain (Norris, 1998) is a random process that incorporates a minimum amount of memory without being totally memoryless, that is a transition from one state to another only depends on the current state. This characteristic is called the Markov property. In a Markov chain, each state corresponds to a discrete observable event, but in the HMM (Rabiner, 1989; Rabiner and Juang, 1993; Huang et al., 2001), the observation itself is a random process. Thus, an HMM is a double-embedded stochastic process consisting of the underlying stochastic process, the Markov chain (state sequence) that is not observable (hidden) but can be observed through another set of stochastic processes associated with each state that produces the sequence of observation features.

An illustration of a 3-state left-to-right HMM is shown in Figure 5.1, in which the state index increases or stays the same with each time step. Generally, left-to-right HMM structures are used to model systems whose properties evolve in a successive manner, such as speech and written language (Rabiner, 1989).

Formally, an HMM is defined by:

- The number of states N and the set of states $\mathbf{Q} = \{q_1, q_2, \ldots, q_N\}$.
- The number of output observation alphabets M and the alphabet itself $\mathbf{O} = \{o_1, o_2, \ldots, o_M\}$[1].
- The state transition probability distribution $\mathbf{A} = \{a_{ij}\}_{i,j=1}^{N}$, where a_{ij} is the probability of a transition occurring from state q_i to q_j.
- The output probability distribution $\mathbf{B} = \{b_i(\mathbf{o})\}_{i=1}^{N}$, where $b_i(\mathbf{o})$ is the probability of emitting an observation $\mathbf{o}$ in state i.
- The initial state probability distribution $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^{N}$.

In conclusion, an HMM is defined by the constants N and M, representing the total number of states and the size of the observation alphabet, the observation alphabet $\mathbf{O}$, and three probability measures $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\pi}$. A compact notation for the set of model parameters for an HMM is represented as

$$\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}). \tag{5.1}$$

There are basically three problems associated with an HMM:

1. **The evaluation problem:** Given an observation sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$ and a model λ, what is $P(\mathbf{O}|\lambda)$, the probability that the model generated the observation sequence?
2. **The decoding problem:** Given the observation sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$ and the model λ, what is the optimal state sequence $\mathbf{Q} = \{q_1, q_2, \ldots, q_T\}$?
3. **The learning problem:** How to adjust the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ to maximise the joint probability $\prod_{\mathbf{O}} P(\mathbf{O}|\lambda)$?

The first problem can be used to evaluate how well a given model matches a given observation sequence. This is especially useful for scoring between different competing models and can be utilised, for example, in pattern recognition. The probability can be calculated from

$$P(\mathbf{O}|\lambda) = \sum_{\forall \mathrm{Q}} P(\mathbf{O}|\mathbf{Q}, \lambda) P(\mathbf{Q}|\lambda). \tag{5.2}$$

[1]Only in the case of a discrete HMM. For continuous density HMM (CD-HMM), single multivariate Gaussian distributions are commonly used for modelling the observation vector.

The direct calculation of $P(\mathbf{O}|\lambda)$ is straightforward, but the number of operations involved is of order of $2TN^T$. Thus, the problem is usually evaluated with the *forward algorithm* (Rabiner, 1989; Huang et al., 2001), which requires only N^2T operations.

The most widely used criterion for finding an optimal state sequence for the decoding problem is to find the single best state sequence that maximises the probability $P(\mathbf{Q}|\mathbf{O},\lambda)$. This can be solved with the Viterbi-algorithm (Viterbi, 1967; Forney, 1973). The decoding problem is used, for example, in ASR to find the best path (optimal state sequence) of letters for the observed acoustic feature sequence.

The third problem, learning, is the most difficult one. No analytical solution is known for solving the model which maximises the probability of the observation sequence. However, iterative algorithms, such as the *Baum–Welch algorithm* (Baum et al., 1970), which utilises the *expectation-maximisation* (EM) algorithm (Dempster et al., 1977), can be used to maximise $\prod_{\mathbf{O}} P(\mathbf{O}|\lambda)$. A speech recogniser or synthesiser involves this learning process to train the HMM with given speech data.

HMMs can be extended with various additional features to make their use more versatile and efficient. For example, null transitions (Bahl et al., 1983), state tying (Bellegarda and Nahamoo, 1990), explicit state duration modelling (Russell and Moore, 1985; Levinson, 1986), and autoregressive HMMs (Poritz, 1982; Juang and Rabiner, 1985; Shannon et al., 2013) are utilised in speech recognition and synthesis. Also, alternative training criteria instead of the common maximum likelihood estimation can be used, such as discriminative training using maximum mutual information (MMI) in speech recognition (Bahl et al., 1986) or minimum generation error (MGE) training in speech synthesis (Wu and Wang, 2006). Other useful features in HMM-based speech synthesis are described in the following parts of this section.

5.2 Speech parameter training and generation using HMM

In HMM-based speech synthesis, the time-varying speech parameters extracted by a vocoder are modelled using left-to-right phoneme HMMs, such as depicted in Figure 5.1. The observation vector consists of continuous-valued speech parameters, and the state-output probabilities are assumed to be single multivariate Gaussian distributions. Thus, the state-

output probabilities are defined as

$$\begin{aligned} b_i(\mathbf{o}_t) &= \mathcal{N}\left(\mathbf{o}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \\ &= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_i|}} \exp\left\{ -\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{o}_t - \boldsymbol{\mu}_i) \right\}, \end{aligned} \tag{5.3}$$

where $\boldsymbol{\mu}_i$ is the d-by-1 mean vector, $\boldsymbol{\Sigma}_i$ is the d-by-d covariance matrix, d is the dimension of the speech parameter vector, and $\mathbf{o}_t$ is an observation vector, representing speech features at frame t.

Let $\mathbf{O} = \left[\mathbf{O}_1^\top, \mathbf{O}_2^\top, \ldots, \mathbf{O}_T^\top\right]^\top$ be the parameters of a speech corpus, W the linguistic specifications extracted from the corresponding text, and $\mathbf{o} = \left[\mathbf{o}_1^\top, \mathbf{o}_2^\top, \ldots, \mathbf{o}_{T'}^\top\right]^\top$ the speech features to be generated according to linguistic specifications w. The training of an HMM system can be written as follows:

$$\lambda_{\max} = \arg\max_{\lambda} p(\mathbf{O}|\lambda, W), \tag{5.4}$$

where

$$p\left(\mathbf{O}|\lambda, W\right) = \sum_{\forall \mathbf{Q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\mathbf{O}_t), \tag{5.5}$$

and $\mathbf{Q} = \{q_1, q_2, \ldots, q_T\}$ is a state sequence. The speech parameter generation, or synthesis, can be written as

$$\begin{aligned} \mathbf{o}_{\max} &= \arg\max_{\mathbf{o}} p(\mathbf{o}|\lambda_{\max}, w) \\ &\approx \arg\max_{\mathbf{o}} \prod_{t=1}^{T'} \mathcal{N}\left(\mathbf{o}_t; \boldsymbol{\mu}_{q_{\max}}, \boldsymbol{\Sigma}_{q_{\max,t}}\right), \end{aligned} \tag{5.6}$$

where

$$\mathbf{q}_{\max} = \arg\max_{\mathbf{q}} P\left(\mathbf{q}|\lambda_{\max}, w\right). \tag{5.7}$$

Equation 5.7 can be maximised using the explicit state-duration probability distributions (Russell and Moore, 1985; Levinson, 1986; Zen et al., 2004), which is discussed in more detail in the following section along with other additional features used in HMM-based speech synthesis.

5.3 Core architecture

An overview of an HMM-based speech synthesis system is shown in Figures 5.2 and 5.3, concentrating on signal flow and statistical modelling with HMMs, respectively. HMM-based synthesis consists of two parts: training and synthesis. As most modern speech synthesisers, HMM-based speech synthesis is a corpus-based method—it requires a recorded speech

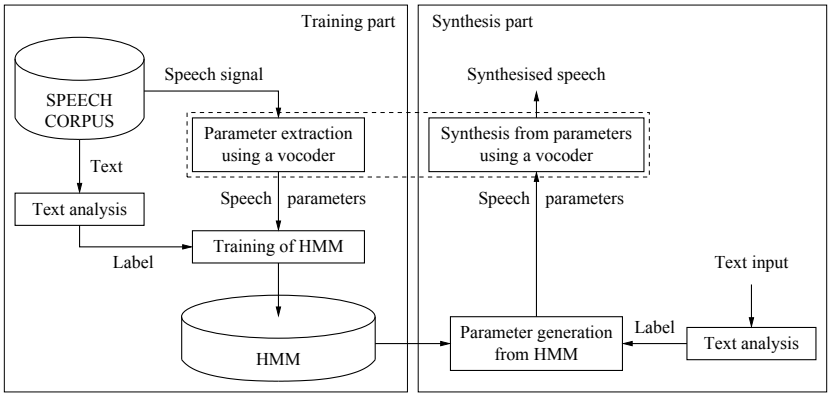


Figure 5.2. Overview of an HMM-based speech synthesis system. The system consists of two stages, training and synthesis. In the training stage, utterances of a speech database are parameterized using a vocoder and trained in a framework of an HMM. In the synthesis stage, speech parameters are generated according to the text input from the HMMs, and speech is synthesised from the parameters using the vocoder.

corpus and the corresponding text as input data. First, in the training part, a vocoder is used to extract speech parameters from the corpus. Also, a linguistic front-end (text analysis) is used to extract phonetic and contextual information from the corresponding text. Various types of vocoders and speech parameterizations can be used, which will be discussed further in Section 5.5. The training of the HMM is performed using the speech parameters and the contextual labels generated by the front-end and aligned according to the speech data. In the synthesis part, the text input is first analysed using the linguistic front-end. Speech parameters are then generated according to the labels and fed to the vocoder, which finally reconstructs a speech signal from the parameters. In the following sections, the specific features that lay the foundation for modern state-of-the-art HMM-based speech synthesis are described in more detail.

5.3.1 Context dependency and parameter tying

As described earlier, the acoustic characteristics of different phonemes are largely context dependent. Thus, the linguistic specifications given by the front-end must be taken into account instead of using simple phoneme specifications. Various linguistic features can be used in HMM-based speech synthesis to describe the context, such as phoneme, syllable, word, and phrase level information as well as lexical stress, pitch accent, tones and break indices (ToBI) (Silverman et al., 1992; Beckman et al., 2005),

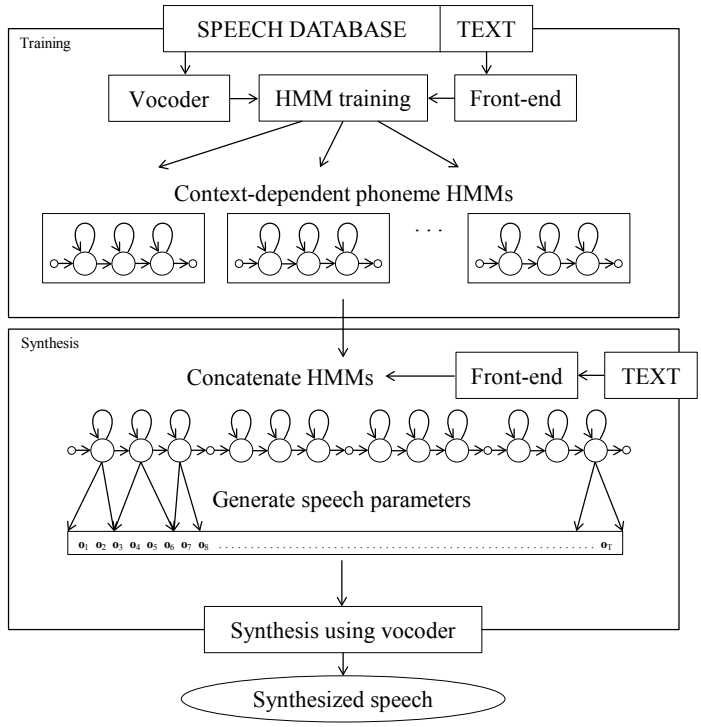


Figure 5.3. Illustration of statistical parametric speech synthesis using HMMs. First, in the training part, speech data is parameterized using a vocoder and the corresponding text is converted into linguistic specifications. Then HMM training is performed to construct context-dependent phoneme HMMs. In the synthesis part, HMMs are concatenated according to linguistic specifications extracted by the front-end from the text input. Speech features are then generated from the concatenated HMM and fed to a vocoder, which generates the synthetic speech waveform.

and POS information, which are all language dependent. For example, the contextual features of the HTS English recipe (HTS, 2014) is shown in Table 5.1. The contextual features are especially important for generating natural prosody and duration for a synthetic utterance.

However, in practice, the amount of speech data is too limited to cover all possible contexts, since the number of different context combinations increases exponentially as the number of contextual factors is increased. The more contextual factors are used, the less data will be used to model each specific context combination, which leads to poor models. To overcome this problem, a state tying approach (Young et al., 1994; Odell, 1995) is used to cluster similar states of the HMMs and to tie the model parameters among several context-dependent HMM states. The state tying is performed by using decision-tree-based context clustering (Yoshimura et al., 1999), which is a top-down, data-driven clustering technique based

Table 5.1. Contextual features of the HTS English recipe (HTS, 2014).

Phoneme	Current phoneme
	Preceding and succeeding two phonemes
	Position of the current phoneme in current syllable
Syllable	Number of phonemes in preceding, current, and succeeding syllables
	Stress and accent of preceding, current, and succeeding syllables
	Position of the current syllable within the current word and phrase
	Number of preceding and succeeding stressed syllables in current phrase
	Number of preceding and succeeding accented syllables in current phrase
	Number of syllables from the previous stressed syllable
	Number of syllables to the next stressed syllable
	Number of syllables from the previous accented syllable
	Number of syllables to the next accented syllable
	Vowel identity within the current syllable
Word	Estimate of the POS of preceding, current, and succeeding words
	Number of syllables within preceding, current, and succeeding words
	Position of the current word within the current phrase
	Number of preceding and succeeding content words in phrase
	Number of words from the previous content word
	Number of words to the next content word
Phrase	Number of syllables in preceding, current, and succeeding phrases
	Position of the current phrase in major phrases
	ToBI endtone of the current phrase
Utterance	Number of syllables, words, and phrases in the utterance

on a greedy algorithm that makes the decision tree grow by splitting the data so as to maximise the likelihood of the data. The idea of decision-tree-based context clustering is shown in Figure 5.4. The size of the decision tree is determined automatically using, for example, the minimum description length (MDL) (Shinoda and Watanabe, 2000) criterion. The decision-tree-based context clustering is performed individually for every stream (e.g., spectral features, f_0, duration, etc.) since they have different context-dependencies. Based on the decision tree, model parameters are then tied across context-dependent HMM states associated with the same class (leaf node), thus being able to represent all possible contexts. In synthesis, appropriate parameters for each state are found by using the decision trees built for each parameter type, which is illustrated in Figure 5.5.

5.3.2 Explicit state duration modelling

In HMM-based speech synthesis, each HMM state has its explicit state duration probability distribution for modelling the temporal properties

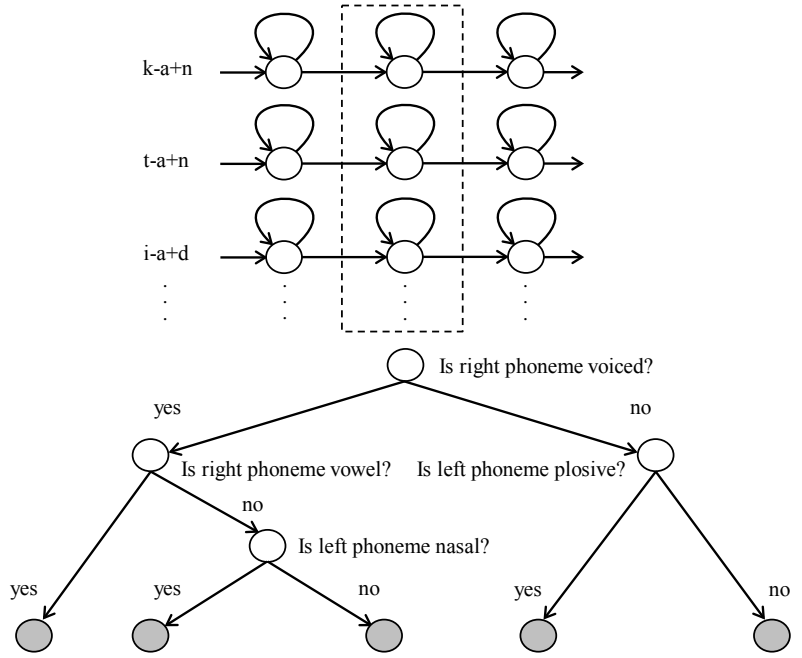


Figure 5.4. Illustration of decision-tree based context clustering for the centre states of phone /a/. The parameters within similar context based on the contextual question set are clustered together and corresponding states in the leaf nodes (marked as grey) are tied in order to avoid poor models due to data sparsity.

of speech (Russell and Moore, 1985; Levinson, 1986; Yoshimura et al., 1998). Although conventional HMM utilises state transition probabilities to determine the duration in each state, it is incapable and too simplistic for modelling speech due to the exponentially decreasing probability with increasing duration. Therefore, hidden semi-Markov models (HSMMs) are used in speech synthesis where Gaussian distributions are used for duration modelling in training and synthesis (Zen et al., 2004). Although, strictly speaking, modern statistical parametric speech synthesisers use HSMMs instead of HMMs, including the explicit state duration modelling, the term HMM-based speech synthesis is often used for convenience.

5.3.3 Incorporating dynamic features

Since each state outputs static mean and variance speech parameter vectors, the output speech parameter trajectories are stepwise sequences instead of having smooth transitions as in natural speech. To overcome this issue, dynamic features (Furui, 1986) are trained along with the static features, usually consisting of first and second time derivatives, which are usually called delta (Δ) and delta-delta (Δ^2) features.

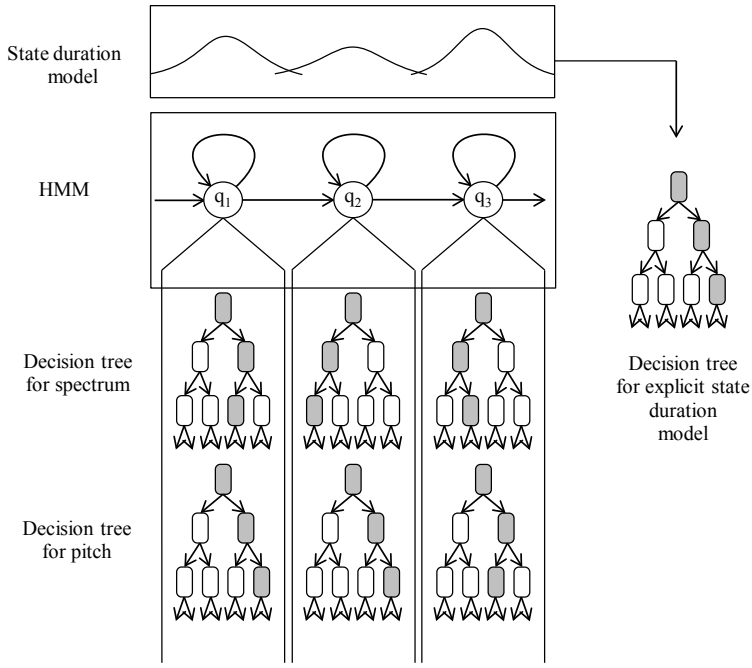


Figure 5.5. Illustration of decision-tree based parameter generation. Each parameter stream (e.g., spectrum, pitch, and duration) has a decision tree which defines the tying of the context-dependent HMM states. In synthesis, the correct tied context-dependent HMM states are determined by the lead node of a decision tree. Hypothetical correct paths in the decision trees are shaded grey.

In parameter generation, the speech parameter trajectories are generated using the Δ and Δ^2 features as constraints, therefore enabling the modelling of the dynamics of speech parameter trajectories (Tokuda et al., 1995a,b, 2000).

5.3.4 Modelling of the fundamental frequency

As the fundamental frequency consists of both continuous f_0 values and discrete symbols[2] indicating unvoiced frames, conventional CD-HMM cannot be used successfully for modelling such discontinuous data. In HMM-based speech synthesis, this problem can be solved, for example, by using multi-space probability distributions (MSD) (Tokuda et al., 1999, 2002a). In MSD-HMM modelling, continuous f_0 values are modelled for voiced frames using a continuous probability distribution, and the voiced/unvoiced decision is modelled by a discrete distribution. By switching between these two distributions according to a label associated with

[2]Unvoiced values are usually indicated by zeros, although its numerical value has no meaning in terms of pitch.

each observation, variable dimensions combining f_0 and voicing information can be modelled in a unified manner. In order to model other speech features as well, like spectral and aperiodicity features, a multi-stream HMM (Young et al., 2006) is used, where each feature set is modelled in individual streams by different output state probability distributions. For example, spectral parameters are modelled by a CD-HMM and f_0 by a MSD-HMM.

Alternatively, the f_0 values can be interpolated for unvoiced regions and then modelled using a CD-HMM (Yu and Young, 2011), while the voicing decisions are made separately. Also, various representations for modelling f_0 have been proposed, such as the discrete cosine transform (DCT) (Teutenberg et al., 2008; Stan and Giurgiu, 2011), the wavelet transform (Suni et al., 2013), or other hierarchical models (Lei et al., 2010), similarly to older studies in f_0 modelling, for example by Öhman (1967) and Fujisaki et al. (1971).

5.3.5 Compensating for over-smoothing

One specific drawback of HMM-based speech synthesis is that the generated speech parameter trajectories are over-smooth compared to the natural ones. This stems from several processes in the HMM training and synthesis. First, the statistical averaging of the parameters for different phonemes in different contexts introduces smoothing—although the averaging improves robustness against data sparsity, the natural variation in the original parameter trajectories cannot be reconstructed after this. Second, the maximum likelihood parameter generation (MLPG) using the dynamic features generates smooth trajectories, thus introducing additional smoothing. The over-smoothing takes effect both in the time and the spectral domain—the variation of time-domain trajectories is reduced and also, for instance, the modelled formants are smoother than natural ones. These effects, for example, make the synthetic speech sound unnaturally buzzy and muffled.

Maybe the simplest method to compensate the over-smoothing is post-filtering, which in this context means emphasising the spectral structure using post-processing after parameter generation. This processing modifies the spectral model so that the dynamics between the formant peaks and the spectral valleys is increased, aiming at a more prominent formant structure depending on the spectral representation. Different post-processing methods have been proposed depending on the spectral repre-

sentation. The adaptive post-filter, originally developed for speech coders (Chen and Gersho, 1995; McCree and Barnwell, 1995), has been adapted for SPSS and implemented in the current HTS version (HTS, 2014; Zen et al., 2007) that generally uses mel-cepstral parameterization (Tokuda et al., 1994). Another technique that works on the line spectral frequency (LSF) domain (Soong and Juang, 1984) was proposed in Ling et al. (2006). A third method was introduced by Raitio et al. (2010) that works on the LP coefficients by appropriately modifying the power spectrum. These methods have been shown to be effective in enhancing the muffled speech quality, although extensive use of post-filtering usually results in an overly sharp formant structure, which results in an artificial speech quality and degrades the similarity of the synthesised speech.

One of the most common methods for preventing over-smoothing is parameter generation using global variance (GV) (Toda and Tokuda, 2007). In this method, speech parameters are generated in a maximum likelihood sense that also considers the original and generated variance of the speech parameter trajectories, thus aiming to reproduce the original dynamics of the speech parameters. Using GV has a similar effect to post-filtering, but also their complementary use is often beneficial. GV can be also implemented in various different ways, such as proposed by Silén et al. (2012). Also, the speech dynamics in the modulation spectrum domain can be compensated in a similar way (Takamichi et al., 2014; Chen et al., 2014), which acts as a frequency dependent GV.

The over-smoothing can also be reduced by utilising natural speech data. Speech parameters can be generated by explicitly using training data for generation (see, e.g., Yu et al., 2007). Alternatively, the original detailed spectral structure can be predicted from a lower level spectral representation (such as mel-cepstrum) using restricted Boltzmann machines (RBM) (Ling et al., 2013), deep belief networks (DBN) (Ling et al., 2013), or DNN (Chen et al., 2014). Also MGE training (Wu and Wang, 2006) may result in a less over-smooth parameter trajectories.

5.4 Flexibility of statistical parametric speech synthesis

In unit selection synthesis, once a voice has been built, there is no easy way to alter the voice characteristics. Voice conversion techniques (Stylianou et al., 1998; Stylianou, 1999) can be used to alter voice characteristics to some extent, but high-quality voice conversion is still prob-

lematic. On the other hand, it is easy to modify and change voice characteristics, speaker identity, speaking style, or emotion in HMM-based speech synthesis by modifying the parameters of the statistical model. Flexibility, among other benefits described in this section, makes SPSS particularly attractive.

Adaptation (mimicking of voices) (Masuko et al., 1997; Tamura et al., 2001; Yamagishi et al., 2009a) is one of the most widely used transformation techniques used in HMM-based speech synthesis. Originally adaptation was used in speech recognisers for adapting the models to a specific speaker or environment to improve recognition accuracy (Gauvain and Lee, 1994; Leggetter and Woodland, 1995). Similar techniques are used in statistical speech synthesis for mimicking, for example, specific voices or speaking styles by using a small amount of corresponding speech material (Masuko et al., 1997; Yamagishi et al., 2009a). Two major techniques have been used for adaptation: maximum a posteriori (MAP) estimation (Gauvain and Lee, 1994) and maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995). The combination of these has been shown to be especially effective (Yamagishi et al., 2009a).

Adaptation techniques have been used for various purposes in HMM-based speech synthesis. First, adaptation can be used to easily create new voices (speaker identities) using average voice-based speech synthesis (AVSS) (Yamagishi et al., 2003; Yamagishi, 2006). In AVSS, an average voice is first trained using a large multi-speaker speech database, after which the model is adapted to a specific speaker using only a few minutes of the target speaker's data (Yamagishi et al., 2008, 2010). In speaker-adaptive training (SAT) (Anastasakos et al., 1996; Yamagishi and Kobayashi, 2007), specific speakers from a large database are used as the initial model for the adaptation, which further increases the adaptation quality. In addition to different speaker identities, also different speaking styles and emotions can be easily reproduced, controlled, and transformed by using similar techniques, such as style modelling (Yamagishi et al., 2005), model adaptation (Tachibana et al., 2006), model interpolation or morphing (Tachibana et al., 2005), and multiple-regression HMMs (Nose et al., 2007). Speaking style adaptation is also used in Publications V, VII, and VIII.

In addition to adaptation, different voices (adapted or separately trained) can be interpolated (Yoshimura et al., 1997) to create intermediate voices by mixing them. Similarly, new voice characteristics can be created using

the *eigenvoice* technique (Kuhn et al., 2000), in which data from multiple speakers is analysed using the principal component analysis (PCA), and the voice characteristics are defined by the combination of eigenvectors (Shichiri et al., 2002). Multiple regression (Fujinaga et al., 2001) has also been used to control voice characteristics more intuitively (Miyanaga et al., 2004; Nose et al., 2007).

SPSS has several other benefits. The footprint of the synthesis module is usually very small, being usually less than 2 MB with the possibility of further reducing the size (Oura et al., 2009). SPSS also has a better coverage of the acoustic space since speech is generated from statistical models. SPSS is robust against various recording conditions (Karhila et al., 2014) or to the lack of some speech units in the case of average voice speech synthesis. SPSS is rather easy to train since there are less tuning parameters than in unit selection synthesis. SPSS is flexible due to the generation of speech by a vocoder, which enables the individual control of the excitation, spectral components, and duration. Finally, multilingual speech synthesis is easier with SPSS since less speech material is required to be recorded. Also, a singing voice synthesis is possible using SPSS (see, e.g., Saino et al., 2006; Oura et al., 2010).

5.5 Vocoders in statistical parametric speech synthesis

In SPSS, the aim is to parameterize speech into a relatively small set of parameters that represents the perceptually relevant characteristics of speech. The parameters should also be suitable for statistical modelling and enable the reconstruction of speech from the parameters. Various types of such parameterizations and analysis/synthesis strategies have been proposed. In this section, different vocoding schemes and parameterization methods used in statistical parametric speech synthesis are presented and discussed.

In speech analysis, a vocoder takes in a speech signal and outputs speech parameters at certain time intervals. Vocoders process speech in short-time frames, that is, the speech signal is segmented to individual frames from which parameters are extracted. Usually the length of a frame is 25 ms, and commonly a frame shift of 5 ms is used, although variations also exist. For instance, to evaluate f_0 for low-pitched male speaker, a longer frame may be required to correctly estimate the lowest f_0 parts.

In synthesis, once the frame-wise parameters of a speech corpus have been modelled using, for example, HMM and parameters are generated from the HMM according the text input, the vocoder reconstructs speech from the parameters. Usually the vocoding process utilises the source-filter model, which is based on the separate modelling of the source and filter parts, although alternative approaches exists, such as the harmonic plus noise model (Erro et al., 2014; Degottex and Erro, 2014) and the dynamic sinusoidal model (Hu et al., 2014). In the following, various methods for estimating and modelling the spectrum and the excitation signal are presented.

5.5.1 Modelling of the speech spectrum

There are two main methods in SPSS for estimating and modelling the spectrum of speech: LP-based methods and cepstrum-based methods.

LP or all-pole modelling is a widely used spectral estimation method (for a review, see, e.g., Makhoul, 1975; Rabiner and Schafer, 1978; Markel and Gray, 1980) that models the resonances (i.e., the formants) of speech using the poles of the LP model. This assumption is acceptable for most speech sounds, but not for nasal sounds or some fricatives. However, by increasing the order of the LP model, also antiformants (i.e., the zeros) can be approximated. Usually the order of the all-pole model is set by

$$p = \frac{f_\mathrm{s}}{1000} + \gamma, \tag{5.8}$$

where f_s is the sampling frequency and γ is set approximately to between 2 and 4. The first part of the equation approximates the number of formants in the modelled frequency range, and the second part (γ) compensates for the poles needed to represent the spectral tilt due to the glottal spectrum. However, in practice, slightly increasing the order will often result in higher speech quality, since a more detailed spectrum will be modelled (Raitio et al., 2011c). It is important to select an appropriate order since too low an order cannot properly model all the formants and too high an order will result in the modelling of the harmonics of the voice source (which is not the goal in estimating the speech spectrum), and thus biased spectral estimates. However, the biasing effect of the voice source harmonics is especially severe with high-pitched speech. In order to alleviate this effect, for example discrete all-pole modelling (DAP) (El-Jaroudi and Makhoul, 1991), weighted linear prediction (WLP) (Ma et al., 1993), true envelope estimation (Roebel and Rodet, 2005; Roebel et al., 2007), or

two-pitch-period synchronous analysis (Babacan et al., 2014) can be used instead. For example, WLP is used in Publication VIII for alleviating the biasing effect of the harmonics in shouted speech.

Alternatively, warped linear prediction, a variant of LP, can be used for the spectral modelling. The idea of warped LP is to use non-uniform frequency resolution, which allows more efficient allocation of the poles with respect to perceptually relevant information in the speech spectrum. Warped LP was originally proposed by Strube (1980), and later it has been studied, for example, in Laine et al. (1994), Tokuda et al. (1994), Karjalainen et al. (1998), and Harma and Laine (2001). For wide-band speech synthesis (Yamagishi and King, 2010), the use of warped LP has been shown to be especially beneficial compared to conventional LP (Raitio et al., 2012b).

LP coefficients are not appropriate for statistical modelling as such, and thus they must be converted into other representation forms before modelling. There are several representations that can be used, of which the most widely used is the LSF representation (Soong and Juang, 1984), which provides good interpolation and smoothing properties needed for the modelling. In synthesis with LP parameters, the direct form filter structure is used. The vocoder presented in this thesis (used in Publications II, III, IV, V, VII, and VIII) uses the LSF-based spectral representation.

The other commonly used spectral representation in SPSS is the cepstrum (Oppenheim and Schafer, 1968). The cepstrum is defined as the inverse Fourier transform (IFT) of the logarithmic magnitude of the Fourier transform of a signal, defined as

$$c(m) = \mathcal{F}^{-1}\big\{ \log |\mathcal{F}\{s(n)\}| \big\}, \tag{5.9}$$

where $s(n)$ is a speech signal with time index n and $c(m)$ is the cepstrum with quefrency index m. Cepstral modelling enables the modelling of both poles and zeros. Usually the mel-cepstrum (Imai, 1983; Tokuda et al., 1994) is used instead of the cepstrum in order to adjust the frequency resolution closer to that of human perception. The cepstrum is truncated to a low order cepstrum in order to only model the overall spectral structure and not the harmonics of the voice source. Usually a cepstral order of 20–60 is used, depending on the modelled frequency range. However, a higher order usually results in improved quality, so special methods may need to be used in order to prevent the biasing effect of the harmonic

peaks. The most common method to prevent this phenomenon is pitch-adaptive time-frequency smoothing (Kawahara et al., 1999), such as used in the STRAIGHT vocoder (Kawahara et al., 1999, 2001), or true envelope estimation (Roebel and Rodet, 2005; Roebel et al., 2007).

Synthesis with mel-cepstral coefficients requires the approximation of the filter with the mel-log spectrum approximation (MLSA) filter technique (Imai, 1983; Fukada et al., 1992). The cepstral-based spectral representation is used in Publications VI and IX.

5.5.2 Modelling of the voice source

One of the key factors for the recent improvements in quality in SPSS has been the advances in the excitation modelling methods. The earliest vocoders used a periodic train of impulses (Makhoul, 1975) located at the GCIs to model the excitation of voiced speech. The quality of impulse-train-excited speech is poor with a buzzy sensation due to the unnaturally identical excitation peaks and the zero-phase character of the excitation. In such a scheme, excitation features other than f_0 and energy cannot be varied, which limits both the quality of speech and the ability to vary voice quality. In order to model the natural aperiodicity in the speech signal, the mixed excitation (Yoshimura et al., 2001) and the two-band excitation (Kim and Hahn, 2007) approaches have been proposed, which mix aperiodic noise with the periodic impulse excitation. In the mixed excitation approach, noise is added to several frequency bands according to aperiodicity weights that define the amplitudes of the periodic excitation relative to the aperiodic noise excitation (see Section 3.2.4 for aperiodicity estimation). Mixed excitation is used, for example, in STRAIGHT (Kawahara et al., 1999, 2001), which is one of the most widely used vocoders in SPSS. In the two-band excitation approach, a maximum voiced frequency is defined above which voiced excitation is composed only of an aperiodic component. Both the mixed and two-band excitation techniques have been shown to improve the synthesis quality compared to systems using the traditional impulse train excitation. In another approach, the closed-loop training method (Maia et al., 2007, 2010), voiced periodic impulse excitation and unvoiced aperiodic noise excitation are fed through state-dependent filters, thus maximising the likelihood of the excitation signal in comparison to the original one. The synthesis quality is greatly improved compared to a conventional impulse train excitation (Maia et al., 2007), and the synthesis quality was comparable to that of a STRAIGHT-

based method (Zen and Toda, 2005). Also parametric models of the glottal flow have been used in speech synthesis (Vincent et al., 2007; Cabral et al., 2007, 2008, 2011b, 2014; Lanchantin et al., 2010; Muthukumar et al., 2013) hence allowing for the ability to modify the voice source characteristics. The results obtained indicate that the problem of buzziness can be partly avoided. In addition, the phase characteristics of the glottal flow have been modelled by Maia et al. (2012, 2013) by using the complex cepstrum. The study shows that the modelling of the phase characteristics results in synthesised waveforms that are closer to natural ones, thus achieving improved speech quality.

The natural excitation of voiced speech, the glottal flow, and its context-dependent variation, is difficult to represent and model using a compressed parametric vector. Therefore, vocoding techniques have been proposed that utilise the excitation waveform *per se* rather than its predefined compressed representation, hence capturing the natural characteristics of the signal, such as the correct phase as discussed in Section 2.5. The excitation signal to be modelled can be either the glottal flow estimated by GIF or the residual computed by LP, for example. The idea of using the natural excitation to improve the synthesis quality is not new (see, e.g., Holmes, 1973; Matsui et al., 1991; Karjalainen et al., 1998; Fries, 1994; Alku et al., 1999), but the development of SPSS and especially vocoders have provided new applications for the approach. In Raitio (2008) and Raitio et al. (2008, 2011c) (last Publication III), a glottal flow pulse estimated from natural speech with GIF is used to construct the voiced excitation. In synthesis, the pulse is first interpolated in time according to f_0, scaled in amplitude based on the energy measure, after which an aperiodic noise component is added to five separate bands in the frequency domain. The pulses are then concatenated in order to create a continuous excitation, which is then modified using an infinite impulse response (IIR) spectral matching filter. The synthesis quality was shown in Raitio et al. (2011c) to outperform STRAIGHT for a low-pitched male voice and to be as good as or better than STRAIGHT in another experiment by Suni et al. (2010) for one male and one female voice. In Drugman et al. (2009b), Drugman and Dutoit (2012), Sung et al. (2010), and Raitio et al. (2013b), PCA is applied to the pitch-synchronous residual/glottal flow signal in order to model the waveform with eigen-residuals, similarly to the methods by Thomas et al. (2009) and Gudnason et al. (2009, 2012). The method in Drugman et al. (2009b) was shown to outperform a simple

excitation, and the method by Sung et al. (2010) was rated better than a simple excitation and a two-band excitation (Kim and Hahn, 2007), and the deterministic plus stochastic model (DSM)-based vocoder (Drugman and Dutoit, 2012) was rated comparable to the quality of STRAIGHT. The study by Raitio et al. (2013b) shows that using the principal components in addition to the mean pulse does not increase the quality of synthetic speech, corroborating the results obtained in the study by Drugman and Dutoit (2012). In Raitio et al. (2014c) (Publication III), it was shown that mean pulses calculated for three different voice qualities (breathy, normal, and Lombard) are also useful for reproducing these voice qualities in synthesis. In Drugman and Raitio (2014) (Publication VI), the impulse excitation, the natural residual, and the eigen-residual were compared using a female and a male voice, arriving at the conclusion that the residuals of natural origin improve the quality for a low-pitched (male) speaker while for a high-pitched speaker the phase information of the natural residual is not of perceptual relevance, and using the natural residual may even deteriorate the quality if the extracted residual is noisy.

Instead of using only a single pulse to construct the excitation, residual/glottal flow pulse codebook based approaches have also been suggested. In Drugman et al. (2009c), a pitch-synchronous residual codebook is constructed, and residual frames are selected to synthesise the excitation. The resulting quality was shown to outperform a simple impulse-train excitation approach. In Raitio et al. (2011a) (Publication IV), a library of various estimated glottal flow pulses is constructed, and pulses are selected for the synthesis of excitation based on a target cost of voice source (and vocal tract) features and a concatenation cost between adjacent pulses. In Raitio et al. (2011a), the pulse library method was shown to be equal in quality to the method in Raitio et al. (2011c) but with slightly better speaker similarity. In Suni et al. (2011, 2012), a pulse library technique was shown to perform comparably to STRAIGHT-based techniques. Raitio et al. (2013b) show that pulse codebook (or library) methods, where individual pulses are selected for synthesis, have a risk of occasionally selecting inappropriate pulses, which can degrade the synthesis quality.

A DNN-based approach for voice source modelling was proposed in Raitio et al. (2014a,b), which avoids the problem of occasionally selecting inappropriate pulses but has the ability to change the excitation wave-

form in response to acoustic speech features. The method uses a DNN to model the context-dependent variability of the glottal flow signal by finding a mapping from acoustic speech features to the sample-wise glottal flow signal. The method in Raitio et al. (2014a) was shown to be equal in quality to a single-natural-pulse-based excitation method (Raitio et al., 2011c), and the method in Raitio et al. (2014b) (Publication V) was shown to provide better voice quality reproduction by synthesising higher quality Lombard speech compared to a PCA-based excitation (Raitio et al., 2013b). However, the speaker similarity was slightly lower compared to the PCA-based method, where a voice-quality-specific pre-selected mean glottal flow pulse was used for each evaluated voice quality. While a speaker-dependent voice source DNN was used in Raitio et al. (2014a,b), a multi-speaker voice source DNN was successfully trained and used for the synthesis of various speakers in Suni et al. (2014).

The DNN-based approach, which predicts the glottal flow signal from various acoustic speech features modelled by an HMM, is well justified since the glottal flow pulse shape and its context dependent variation is indeed dependent on the acoustic features. The acoustic features in the method include the vocal tract spectrum, the voice source spectrum, f_0, energy, and the HNR. These speech features and the glottal flow shape are speaker dependent in general (Fant, 1997). The glottal flow shape also varies according to f_0 (Strik and Boves, 1992; Tooher and McKenna, 2003; Fant, 1997), the phonetic context (Tooher and McKenna, 2003; Fant, 1997), prosody (Strik and Boves, 1992; Fant, 1997; Airas et al., 2007; Vainio et al., 2010), and voice quality (Gobl and Ní Chasaide, 2003), which are reflected in the acoustic features (Lorenzo-Trueba et al., 2012). Also modelling the source-filter interaction[3] is possible, at least in theory, since the vocal tract spectrum is used as an input feature to predict an appropriate glottal flow pulse. However, the pseudo-random variation from pulse to pulse that occurs in natural speech and may be crucial for the naturalness of speech, cannot be modelled with the current DNN-based method

[3]The source-filter interaction can be interpreted in two slightly different ways. Conventionally, the source-filter interaction means effects such as skewing, ripple, and damping of the glottal flow pulse due to the interaction between the glottal flow and the vocal tract (Rothenberg, 1981; Ananthapadmanabha and Fant, 1982; Ananthapadmanabha, 1984; Fant et al., 1985b; Lin, 1987; Fant and Lin, 1987; Klatt and Klatt, 1990; Teager and Teager, 1990). In speech synthesis, however, the source-filter interaction may also mean that the modelling of these two components should not be considered independent of each other (Merritt et al., 2014; Henter et al., 2014).

since the input acoustic features vary smoothly and, therefore, so does the predicted glottal flow shape.

Although natural glottal flow pulse based methods, such as the DNN-based method, produce an improved model of the glottal flow signal regarding correct shape and phase characteristics, the modelling of the aperiodic component still remains a challenge. Various methods to measure aperiodicity have been proposed (basic techniques were reviewed in Section 3.2.4). A simple and robust technique for noise spectral weighting makes use of a maximum voiced frequency F_{m}, which depicts a boundary between the periodic low-frequency component and the aperiodic high-frequency component. This method is commonly used to model the excitation in SPSS (Drugman et al., 2009b; Drugman and Dutoit, 2012; Lanchantin et al., 2010). An improved version of the aperiodicity estimation using F_{m} was presented by Drugman and Stylianou (2014). Another technique is to estimate the amount of aperiodicity in each spectral band. For example, in mixed excitation (Yoshimura et al., 2001), the voicing strength in each band is estimated using the normalised correlation coefficient around pitch lag. In Raitio et al. (2008, 2011c), the strength of the cepstral peak at the pitch lag is used to measure the periodicity at each band. In Kawahara et al. (2001) and Raitio et al. (2011a), aperiodicity, or HNR, is determined based on the ratio between the upper and lower smoothed spectral envelopes and averaged across frequency bands according to the ERB scale (Moore and Glasberg, 1983). Such a technique is used both in the STRAIGHT (Kawahara et al., 1999, 2001) and the latest GlottHMM (Raitio et al., 2011c,a, 2014b) vocoders. The perceptual effect of different aperiodicity measurement techniques and methods of mixed noise with the periodic component has been investigated in the study by Drugman and Raitio (2014), which is the Publication VI of this thesis. The study shows that using a noise model is essential in improving the synthesis quality, and the dynamic modelling of F_{m} (Drugman and Dutoit, 2012) and the bandwise HNR-based model (Raitio et al., 2011a) are both appropriate for synthesising the aperiodic component. On the other hand, the perceptual impact of the noise time-envelope, that is, the distribution of noise energy in time per pitch period, seems to be negligible.

5.6 Future directions

SPSS has developed at a very fast pace during the past ten years, and it has clearly exceeded the popularity of unit selection methods in research. The technology originally used for speech recognition has been successfully used for automatic speech generation, which has sped up the progress. Recently, the two synthesis paradigms have converged towards each other, which has led to hybrid techniques combining the benefits of SPSS and unit selection synthesis. There are still fundamental restrictions in both methods, but combining the naturalness of unit selection synthesis and the flexibility of SPSS is a reasonable goal in the near future.

Another new paradigm in SPSS is the introduction of deep learning. Previously the training of deep architectures of artificial neural networks was widely considered too problematic, but new algorithms, increased computing power, and large corpora have produced remarkable results in speech recognition (Hinton et al., 2012), and now similar techniques are used for speech synthesis with promising results (Ling et al., 2013; Kang et al., 2013; Zen et al., 2013; Fernandez et al., 2013; Lu et al., 2013; Zen and Senior, 2014; Fan et al., 2014a,b). There are several limitations in the current decision-tree clustered context-dependent HMM approach. First, the spectrum is hard to model directly due to high dimensionality and strong correlation between adjacent spectral bins. Using the spectrum instead of its compressed representation (e.g., LSF or mel-cepstrum) has shown to yield better synthesis quality (Ling et al., 2013; Chen et al., 2014). Second, data fragmentation occurs using the decision-tree context clustering, and thus it is inefficient for representing complex dependencies between linguistic and acoustic features. On the other hand, deep learning can efficiently model high-dimensional, highly correlated features, such as speech spectrum, and it can automatically integrate the feature extraction and acoustic modelling. Deep learning methods are also exponentially more efficient than fragmented methods, such as decision-tree-based context clustering (Young et al., 1994; Odell, 1995). Current deep learning methods, such as the deep belief network (DBN) (Ling et al., 2013; Kang et al., 2013), the DNN (Zen et al., 2013; Lu et al., 2013; Fan et al., 2014a), the DNN-Gaussian process (Fernandez et al., 2013), the mixture density network (MDN) (Zen and Senior, 2014), and the long short-term memory (LSTM) recurrent neural network (RNN)

(Fan et al., 2014b) have already given promising results in combination with HMMs or without. Generally, deep learning is prone to overfitting, and thus large amounts of data may be required for successful training. Although some studies have used huge amounts of speech material (Zen et al., 2013; Zen and Senior, 2014), others have achieved successful results even with moderate-sized data (Fan et al., 2014a,b). The deep learning scheme is developing fast, and probably new methods will find applications in speech synthesis, which will lead to improved synthesis quality and possibly also increased flexibility. The flexibility of deep learning is still questionable since the same adaptation methods as with HMM cannot be used. There are methods for adapting DNNs (see, e.g., Yao et al., 2012; Saon et al., 2013; Liao, 2013; Yu et al., 2013a; Deng et al., 2013; Swietojanski and Renals, 2014), however, they have not yet been applied to speech synthesis.

In the near future, deep learning will be used increasingly for speech synthesis. Methods that integrate feature extraction and mapping from linguistic features to acoustic ones will be used. Possibly only speech frames and corresponding text information will be used as an input to a deep learning architecture, which finds the optimal feature extraction and linguistic features, and can thus generate new speech frames based on new input text. The first steps in this direction have already been taken is ASR, as the method in Tüske et al. (2014) takes a raw speech signal as an input. Also speech synthesis is being approached with direct waveform modelling (Vishnubhotla et al., 2010; Raitio et al., 2014b,a) using deep learning architectures (Hinton and Salakhutdinov, 2006; Hinton et al., 2012) and with unsupervised learning of linguistic features from text (Watts, 2012; Lu et al., 2013).

TTS synthesis methods that utilise the source-filter model may be improved by modelling the two components, source and filter, in a unified manner. Currently, the source and filter are modelled independently, neglecting the source-filter interaction and the fact that separating the two components is a difficult task, and usually it is performed in a rather arbitrary manner. The degrading effect of this assumption has been observed and argued in recent speech synthesis studies by Merritt et al. (2014) and Henter et al. (2014). The current excitation methods are unable to model the source-filter interaction, although some methods, such as the ones by Raitio et al. (2014a,b), have the capability to model this phenomenon in theory. However, the method in Raitio et al. (2014a,b) is unable to

model longer context inter-pulse variations, which might be needed for improved quality. Deep learning may also provide a solution to this problem, since longer context can be easily modelled, for example, by using stacked frames (Chen et al., 2014) or recurrent architectures such as the LSTM (Fan et al., 2014b). However, such approaches have not yet been tested extensively. For example, the study by Fan et al. (2014b) seems to improve especially the modelling of the long-term prosodic information but less the segmental speech quality. This is probably due to the source-filter model utilised in the study, which neglects interaction. Also, using direct waveform modelling instead of a source-filter-model-based vocoding should, in theory, provide an adequate model of the source-filter interaction.

5.7 Summary

This section presented the basic theory and methods used in SPSS. HMMs were first shortly reviewed, after which the speech parameter training and generation using HMMs was described. The core architecture and special methods used in HMM-based speech synthesis were also presented. The flexibility and other benefits that statistical parametric synthesis offers were described, after which various vocoder technologies, emphasising the spectrum and voice source modelling methods, were presented. Finally, possible future directions of SPSS were provided with the main focus on deep learning methods.

6. Summary of publications

This section summarises the publications in the thesis.

Publication I: "Automatic glottal inverse filtering with the Markov chain Monte Carlo method"

In the first publication (Auvinen et al., 2014), a new GIF method is proposed that makes use of a simple glottal flow model, the Rosenberg–Klatt (RK) model (Rosenberg, 1971; Klatt, 1980; Klatt and Klatt, 1990), and Bayesian inversion (Kaipio and Somersalo, 2005) using the Markov chain Monte Carlo (MCMC) sampling method (see, e.g., Gilks et al., 1996; Hastings, 1970; Gamerman, 1997; Smith and Roberts, 1993; Tierney, 1994; Roberts and Smith, 1994). The new method first estimates an initial vocal tract model and a glottal flow signal using an existing inverse filtering method, the IAIF (Alku, 1992). Then, the open phase of the glottal flow model and the radii and angles of the first eight poles, defining the four first formants, are varied using MCMC. A new signal is synthesised using the new vocal tract model and the glottal flow model, which is then compared to the original speech frame to get feedback to the MCMC estimation, which aims to minimise the error between the synthetic and the original speech waveforms. MCMC approximates the posterior distribution of the parameters, and the final estimate of the vocal tract is found by averaging the parameter values of the Markov chain. By adjusting the poles of the initial vocal tract and the glottal flow models, a more accurate vocal tract estimate is obtained, which is less affected by the biasing effect of the voice source harmonics.

The proposed method, MCMC-GIF, is compared with two well-known GIF methods, IAIF (Alku, 1992) and the complex-cepstrum-based decomposition (CCD) (Drugman et al., 2009a, 2011). Since the reference glot-

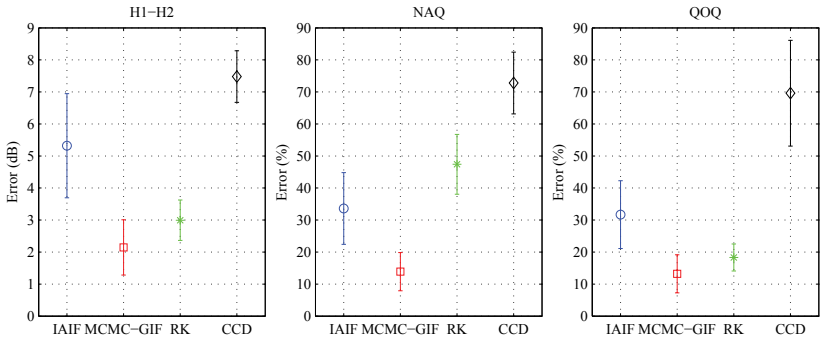


Figure 6.1. Overall average error of H1–H2, NAQ, and QOQ for IAIF, MCMC-GIF, the fitted RK model, and CCD in GIF of synthetic vowels. Data are represented as means and 95% confidence intervals.

tal flow waveform is unknown in natural speech, synthetic vowel data generated using the physical modelling of the vocal folds and the vocal tract (Alku et al., 2006b, 2009) was used in the experiments. The inverse-filtering performance was measured using H1–H2 (Titze and Sundberg, 1992), the normalised amplitude quotient (NAQ) (Alku et al., 2002), and the quasi-open quotient (Hacki, 1989). The summarised results with synthetic speech are shown in Figure 6.1.

Overall, the results show that the proposed method gives more accurate inverse-filtering results compared to the two well-known reference methods. Examples with natural speech also show positive results for the proposed method. The need for accurate glottal closure instant detection and high computational load are the drawbacks of the proposed method. Nevertheless, the study shows that the proposed method is feasible and further developments can be made for practical applications.

Publication II: "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction"

In the second article (Airaksinen et al., 2014), another new glottal inverse filtering method is proposed. The method is based on performing closed phase analysis (Strube, 1974; Wong et al., 1979) over multiple pitch periods using WLP (Ma et al., 1993) with a special weighting function called the attenuated main excitation (AME). The weighting window downgrades the contribution of the glottal excitation in the LP model optimisation. The new method is thus called quasi-closed phase inverse filtering (QCP).

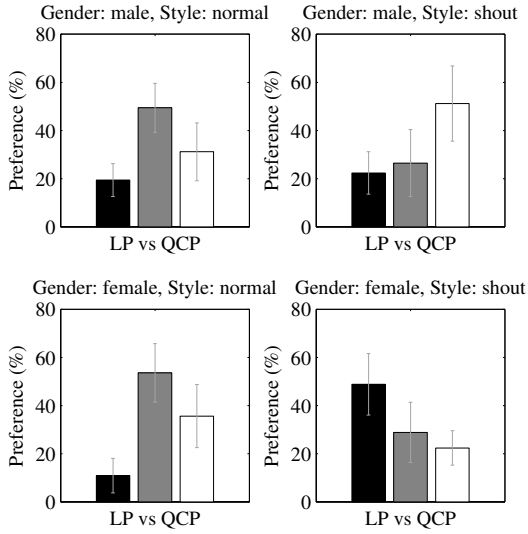


Figure 6.2. Subjective listening test results comparing the LP and QCP methods in vocoder analysis-synthesis quality for different voice types. Data are presented as means and 95% confidence intervals. The grey bars indicate the results for no preference.

The new QCP method is compared to four other GIF methods: CP (Strube, 1974; Wong et al., 1979), IAIF (Alku, 1992), CCD (Drugman et al., 2009a), and WLP (Ma et al., 1993) using the short-term energy weighting function. The test data includes both synthetic vowels produced with the conventional source-filter model using the LF glottal flow model (Fant et al., 1985a; Fant, 1995) as an excitation as well as data produced using the physical modeling approach (Alku et al., 2006b, 2009). Various glottal flow parameterization methods are used to measure the performance of the GIF methods, namely, NAQ (Alku et al., 2002), QOQ (Hacki, 1989), H1–H2 (Titze and Sundberg, 1992), HRF (Childers and Lee, 1991), and the mean-squared error of the estimated glottal flow waveform. The results show that the new QCP method improve the GIF performance both in low and high-pitched speech. In addition, the QCP method is used within a physiologically oriented vocoder (presented in Publication III) to evaluate the analysis-synthesis quality of speech and shout. The subjective evaluations show that by using the new QCP method instead of the IAIF method improved the perceptual quality of the vocoded normal and shouted speech. Figure 6.2 shows the results of the subjective listening test comparing the LP and QCP methods in vocoder analysis-synthesis quality for different voice types.

Publication III: "HMM-based speech synthesis utilizing glottal inverse filtering"

The third article (Raitio et al., 2011c) presents a new vocoder that utilises GIF in speech parameterization. The vocoder, called GlottHMM, utilises the IAIF GIF method (Alku, 1992) to decompose speech into a voice source signal and a model of the vocal tract, thereby enabling the detailed analysis, parameterization, and modelling of the voice source signal and the vocal tract spectrum. The voice source signal is parameterized into several features, namely, the fundamental frequency, the HNR, and voice source spectrum. In synthesis, the excitation voice source signal is reconstructed by modifying and concatenating a pre-computed glottal flow pulse. The pulse is first interpolated in time and scaled in magnitude to match the given fundamental frequency and energy, respectively, after which noise is added in the spectral domain based on the band-wise HNR measure to produce the correct degree of voicing. Finally, the spectrum of the excitation is modified using a spectral matching IIR filter in order to control the spectral tilt and also the spectral details of the voice source. The synthesised excitation is finally filtered with the vocal tract filter to create speech. This vocoder scheme enables the more accurate reconstruction of the voice source and preservation of the glottal flow signal phase. The parameterization also enables the separate modelling and modification of the voice source and the vocal tract filter.

The proposed vocoder is compared with two other commonly used vocoders, one with an impulse-train-based excitation with mel-cepstral spectral modelling (Imai, 1983) and the most widely used vocoder STRAIGHT, which uses a mixed excitation scheme (Yoshimura et al., 2001) and mel-cepstral spectral modelling as well. The proposed method outperformed both methods in terms of synthesis quality and intelligibility with a Finnish male voice. Figure 6.3 shows the results of a CCR subjective listening test comparing all the three methods, indicating that the proposed method achieves the best synthesis quality. Figure 6.4 shows the results of a pair comparison test between the proposed method and STRAIGHT, which indicate that the proposed method is almost always preferred over the STRAIGHT method.

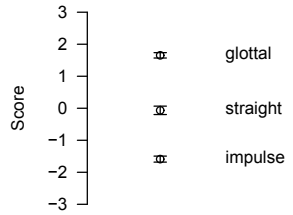


Figure 6.3. Results of the subjective listening test comparing the following systems: proposed system (glottal), STRAIGHT-based system (straight), and impulse-train excited system (impulse). The mean score has no explicit meaning, but the distances between the scores define the amount of preference relative to each other. The 95% confidence intervals are presented for each score.

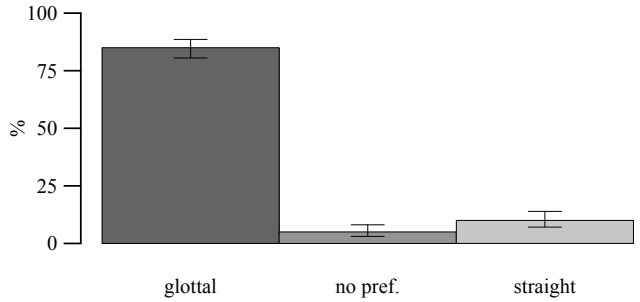


Figure 6.4. Results of the pair comparison test applied to the proposed system (glottal) and the STRAIGHT-based system (straight). The bars indicate the percentage of the total number of answers to the question "Which one would you rather listen to?". The centre bar (no pref.) indicates no preference for either of the methods. The 95% confidence intervals are presented for each bar.

Publication IV: "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis"

This conference paper (Raitio et al., 2011a) extends the work presented in Publication III by introducing a glottal flow pulse library for the excitation generation, in addition to other important refinements to the GlottHMM vocoder. The main difference between the original vocoder proposed in Raitio et al. (2011c) and this work is the excitation generation technique. While in the original implementation, a single glottal flow pulse is used and modified for excitation generation per utterance, the new method utilises a library (or codebook) of various glottal flow pulses, from which the best matching pulses are selected at synthesis time for excitation generation.

In addition to the methods in the original vocoder, such as GIF, the new algorithm requires GCI detection in order to extract individual glot-

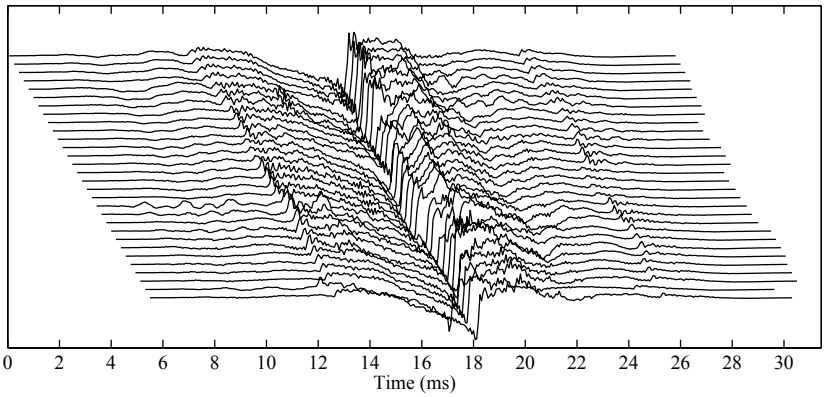


Figure 6.5. Windowed two-period glottal volume-velocity derivative waveforms from the pulse library of a male speaker extracted with the automatic analysis method.

tal flow pulses from the voice source signal. After GCI detection, each two-pitch period glottal flow derivative pulse is extracted and windowed with the Hann window, and stored in a library. Figure 6.5 shows a set of these windowed glottal flow derivative waveforms. The speech parameters extracted by the vocoder are linked with each pulse in the library. Thus, at synthesis time, a search is performed to find the best matching pulses from the library based on the target cost of the speech features and the concatenation cost of adjacent pulses. This method enables the generation of a rich and varying voice source signal for synthesising more natural speech with specific speaker characteristics and different voice qualities. Also, a new HNR estimation technique is introduced. While the old method used cepstral peak prominence for the task, the new method evaluates the degree of voicing from the glottal source signal based on the ratio between the upper and lower smoothed spectral envelopes (defined by the harmonic peaks and interharmonic valleys, respectively), which are then averaged across five frequency bands according to the ERB scale (Moore and Glasberg, 1983).

The new method is evaluated in a series of subjective listening tests. First, the quality difference between the existing and the new method is evaluated using a CCR listening test, where the two methods show similar performance—the new method is rated slightly higher in quality, but the result is not statistically significant. In a second test, natural speech is also included in the CCR test. The results are similar to the first test: the two methods show similar performance, but both are rated less natural than natural speech. Finally, a speaker similarity test is performed to

find out if either of the methods can reproduce the speaker characteristics more accurately. In the similarity test, listeners are presented with two samples at a time (one from each method) and asked which one of the two samples sounds more similar to the speaker in the third reference sample. The results of the similarity test show that the new pulse-library method is able to better reproduce speaker characteristics in comparison to the existing method.

The new method has the benefit of being able to generate a more natural excitation signal by imitating the natural variation in the voice source, but it also suffers from the difficulty of selecting appropriate pulses, which may distort the synthetic speech if not successful.

Publication V: "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort"

The second conference paper (Raitio et al., 2014b) introduces a new excitation generation technique based on DNN. The new method is based on learning a mapping between the acoustic speech features extracted by the vocoder and the time-domain pitch-synchronous glottal flow waveform. The DNN-based voice source modelling method partly solves the issues with the methods presented in Publications III and IV, that is, it achieves appropriate variation in the glottal flow pulse shape in response to the acoustic speech features, but it avoids the occasional errors when selecting pulses from a pulse library.

The study shows how the proposed DNN-based voice source modelling method can be used to synthesise various degrees of vocal effort (from breathy through normal to Lombard) using only a single DNN trained from speech data including the three different vocal effort levels. The DNN-based method is compared to a reference method where a PCA-based pulse is manually selected for each speaking style to account for the required voice quality change. A demonstration of the pulse modelling capability is illustrated in Figure 6.6, where pulses are generated for various degrees of vocal effort using the corresponding adapted acoustic speech features.

Subjective evaluations show that the proposed method is equal in quality to the manual method with breathy and normal speaking styles, and better with Lombard style. However, speaker similarity is rated worse with breathy speech. This most probably stems from the spectral match-

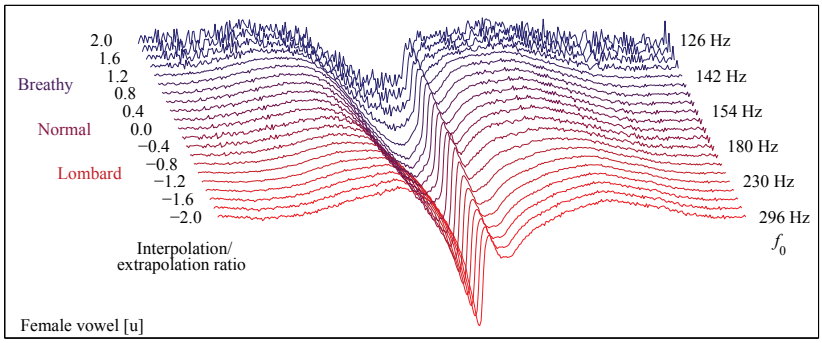


Figure 6.6. Demonstration of the DNN-based excitation modelling by interpolating and extrapolating different HMM-based speaking styles from original breathy (1.0), normal (0.0), and Lombard speech (−1.0) and generating the DNN-based pulses corresponding to the generated speech parameters of various degrees of the styles. The resulting pulses (without interpolation in time, scaling in magnitude or noise added) are shown for vowel [u] uttered by a female speaker.

ing scheme that is not used in the current DNN-based method for simplicity. However, such a scheme can also be used with the proposed DNN-based method if desired.

In summary, the paper shows that the proposed DNN-based voice source modelling method is capable of successfully reproducing different degrees of vocal effort with high quality. Moreover, the method is completely automatic, and thus no manual tuning is required, as was done in the reference method. The proposed method is also more robust compared to previous methods, such as the pulse-library-based method (Raitio et al., 2011a), which is prone to errors in the pulse selection.

Publication VI: "Excitation modeling for HMM-based speech synthesis: Breaking down the impact of periodic and aperiodic components"

In the third conference paper (Drugman and Raitio, 2014), the perceptual effects of the accurate modelling of the periodic and aperiodic components in voice source modelling are investigated using a generalised mixed excitation. There are three main components that can affect the quality of synthesis: 1) the type of periodic waveform used to synthesise voiced speech, 2) the spectral weighting of the aperiodic component in voiced speech, and 3) the time envelope of the aperiodic component. The aim of the study is to evaluate the relative perceptual importance of each factor and to seek the most appropriate method to model the periodic and

aperiodic components.

A generalised mixed excitation scheme is used to study all the three components. Three types of periodic excitation schemes are used: impulse excitation, eigenresidual (Drugman and Dutoit, 2012), and a residual extracted from natural speech. Four types of noise spectral weighting methods are studied: no added noise at all, noise added according to the fixed maximum voiced frequency (F_{m}) (Stylianou, 2001), noise added according to a dynamic F_{m}, and noise added according to the HNR measure used in Raitio et al. (2011a). Also, three types of noise time envelopes are evaluated: constant, triangular, and DSM-based (Drugman and Dutoit, 2012) time envelopes.

A large subjective listening test is conducted using the different methods and a female and a male English speech database. Since it is not practical to test all possible combinations of the different methods, a sequential evaluation method is adopted. First, the noise spectral weighting methods are evaluated using the simplest impulse excitation. Then, the effect of noise time modulation is evaluated using the impulse excitation and the highest rated noise spectral weighting. Finally, the three periodic waveforms are evaluated using the highest rated noise time envelope and noise spectral weighting methods revealed by the first two evaluations.

The results of the first test, depicted in Figure 6.7 (top graph), show that the spectral weighting methods HNR and dynamic F_{m} are rated equal for male speech, while dynamic F_{m} is rated better for female speech. The fixed F_{m} method and a pure impulse excitation without noise are always rated the worst. The results indicate that adding noise to voiced excitation is beneficial and HNR and dynamic F_{m} are both appropriate methods for the spectral weighting. Since the system with dynamic F_{m} is rated higher in quality for female speech, it is chosen as the spectral weighting method for the following evaluations.

The results of the second test, depicted in Figure 6.7 (middle graph), show that the noise time envelope does not have a perceptually relevant effect, and thus the simplest one, the constant time envelope, is chosen for the third evaluation.

The results of the third test, depicted in Figure 6.7 (bottom graph), show that results diverge for male and female speakers. For the male voice, the natural residual frame and the eigenresidual have the highest quality ratings while the impulse excitation is rated lower in quality than the natural residual. For the female speaker, impulse excitation and eigenresidual

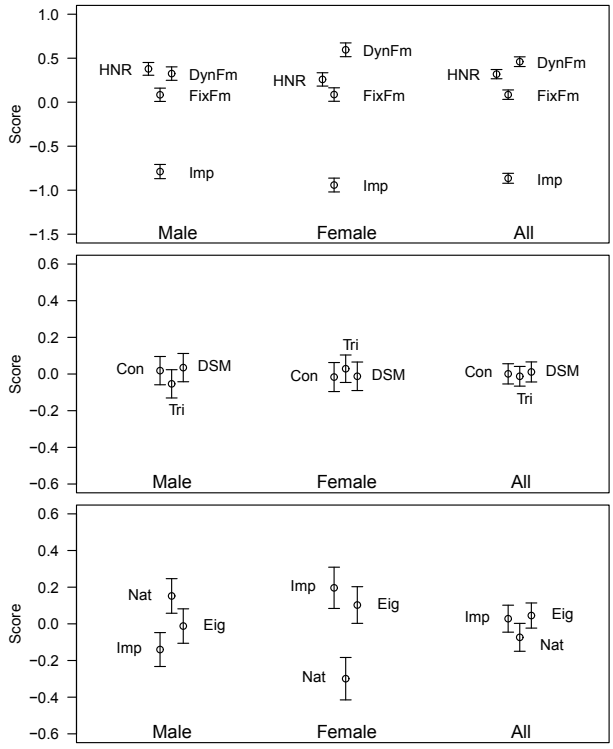


Figure 6.7. Results (mean and 95% confidence intervals) of the subjective evaluation comparing noise spectral weighting (uppermost), the noise time envelope (middle), and the periodic waveform (bottom).

are rated equal in quality while natural residual is rated the worst. The results indicate that the perception of the periodic waveform depends on the f_0 of the speaker. This is due to the human phase perception where the phase of a signal with a low repetition rate (pitch) has a perceptual effect which vanishes when the repetition rate is increased enough. Thus, with male speech, the natural residual is perceived as the most natural one, but in the case of female speech, impulse excitation and eigenresidual give the best results, although impulse excitation completely lacks the natural phase of the original excitation. The natural excitation is rated significantly lower in quality due to the (undesired) noise present in the female residual pulse, which results in a buzzy sound quality.

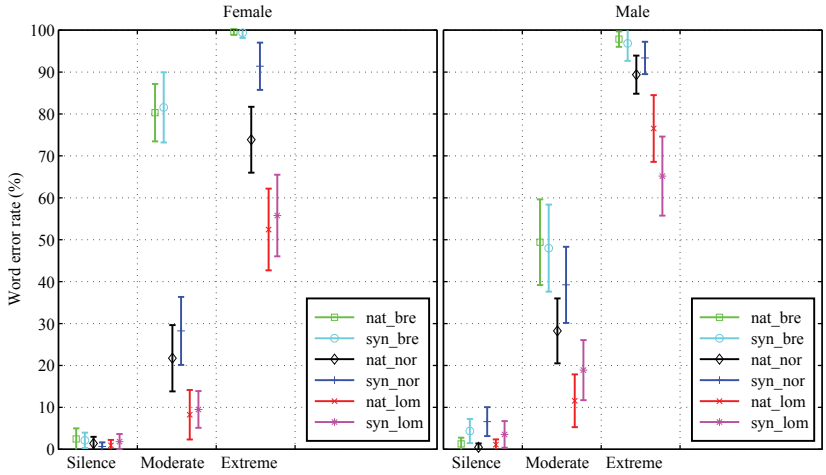


Figure 6.8. Results of the intelligibility test for the female (left) and male (right) voices in three noise conditions: silence, moderate street noise (63 dB, SNR = −1 dB), and extreme street noise (70 dB, SNR = −8 dB).

Publication VII: "Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise"

The fourth journal article (Raitio et al., 2014c) studies the synthesis of speech with varying vocal effort and its perception in various noise environments. Three types of speech are recorded from a male and female subjects, breathy, normal, and Lombard speech, and the corresponding synthetic voices are built using the vocoder presented in Publications III and IV with additional modifications. The voice building consists of first computing the mean glottal flow pulses for each speaker and each vocal effort level from the pulse libraries extracted from the corresponding speech corpora. Then, normal (modal) speaking style voices are trained using the normal speech material for each speaker. In the synthesis of varying vocal effort, normal voice models are first adapted to breathy or Lombard speech using the corresponding breathy or Lombard speech data, and the mean pulses of the corresponding vocal effort level are used to reconstruct the excitation.

The intelligibility, quality, and suitability of the natural and synthetic (loudness normalised) samples of breathy, normal and Lombard speech are evaluated in three types of realistic multichannel noise environments: silence, moderate street noise (63 dB), and extreme street noise (70 dB). The results of the intelligibility, shown in Figure 6.8, show that increased vocal effort improves the intelligibility of speech both for natural and synthetic voices. Although the synthetic voices generally have slightly higher

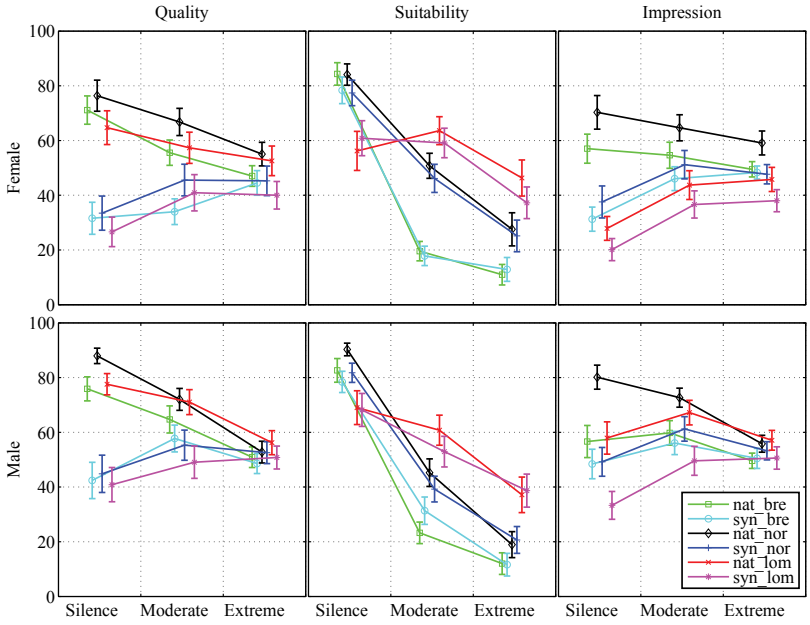


Figure 6.9. Results of the subjective evaluation for female (upper) and male (lower) voices. The measured quantities are quality, suitability, and impression of the speaking style (pleasant vs. irritating).

word error rates than natural speech, the reproduction of vocal effort in synthesis is successful.

The results of the subjective evaluation of quality, suitability, and impression of the speaking style are shown in Figure 6.9. The results show that the synthesised voices with varying vocal effort are rated very similarly to their natural counterparts. Especially the suitability ratings of natural and synthetic voices have a very high correlation. Only the quality ratings show a clear separation between the natural and synthetic voices, but the perceived quality differences decrease or vanish as the SNR is decreased.

The results indicate that when synthetic speech is reproduced in natural environments with background noise, a speaking style adaptation is beneficial both in terms of intelligibility and suitability. Especially Lombard speech is evaluated as more intelligible and suitable to be heard in the presence of noise. However, the breathy speaking style is not considered more appropriate nor more pleasant in the silence than normal speech. The study show that the quality of synthetic speech is not significantly degraded by the adaptation, and in the presence of noise, the degradation of speech quality caused by statistical modeling and vocoding loses its significance, therefore justifying the use of such synthetic voices in the

presence of noise.

Publication VIII: "Analysis and synthesis of shouted speech"

In the fourth conference paper (Raitio et al., 2013c), the acoustic properties of shouted speech are analysed in relation to modal speech, and various techniques are studied for generating synthetic shouted speech. The analysis of the acoustic parameters of normal and shouted speech, collected from 12 male and 12 female Finnish speakers, show large differences between the two styles, which induces difficulties in conventional speech processing methods. Especially the high f_0 of shouted speech may distort the spectral estimates of speech. Due to this, three different spectral estimation methods are compared in analysis-synthesis experiments using the vocoder presented in Publications III and IV. The three methods are LP, WLP with an STE window (Ma et al., 1993), and WLP with an AME window (Alku et al., 2013) (similar to the method in Publication II). The results of a subjective evaluation using the analysis-synthesis samples show that the WLP with the AME window is best suited for modal speech and WLP with the STE window is best for shouted speech. The results show that using a spectral estimation method that is not biased by the sparse harmonics of the voice source is beneficial.

The synthesis of shouted speech is performed using two different techniques, through adaptation and voice conversion. In the adaptation experiment, two spectral estimation techniques were compared: LP and WLP with the STE window. The results show that the method using WLP with the STE window gives s better adaptation quality for both the male and female voices.

Finally, natural and synthetic normal and shouted speech are assessed in a subjective listening test in order to evaluate the quality, degree of perceived shouting, and degree of perceived vocal effort. The results of the evaluation are shown in Figure 6.10. Both voice conversion-based and adapted shouted speech are evaluated. The results show that the synthetic shouting voices preserve the impression of shouting and the used vocal effort fairly well, although the quality is degraded due to processing in adaptation and voice conversion. The adaptation and voice conversion techniques are rated equal in quality, but the degree of shouting and perceived vocal effort is better preserved with the adaptation method.

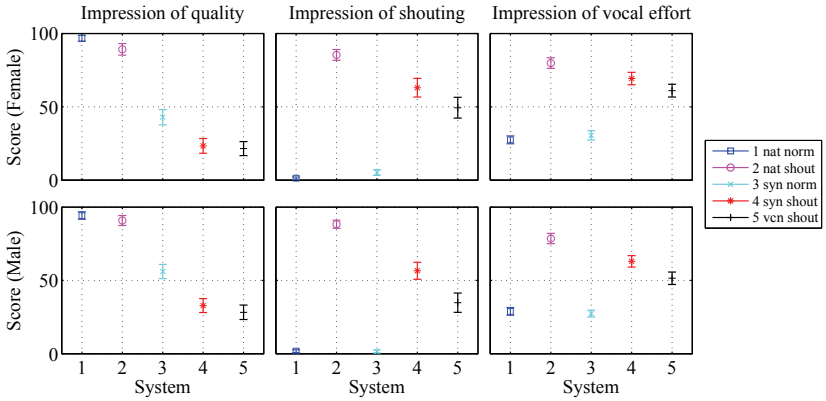


Figure 6.10. Results of the subjective evaluation (impression of quality, shouting, and used vocal effort) for natural and synthetic normal and shouted male and female voices. Synthetic shouted voices were created by adaptation (syn shout) and voice conversion (vcn shout).

Publication IX: "HMM-based synthesis of creaky voice"

The fifth and last conference paper of this thesis (Raitio et al., 2013a) studies the synthesis of creaky voice. Creaky voice quality is frequently used in many languages, and therefore the synthesis of creaky voice was expected to enhance the naturalness of synthetic speech if present. The study is part of previous work by the authors, including, for example, creaky voice detection (Kane et al., 2013b), creaky excitation modelling (Drugman et al., 2012b), and prediction of creaky voice from context (Drugman et al., 2013). In this paper, a fully functioning creaky voice synthesis is built. The method is based on first detecting creaky voice from a creaky speech corpus using an existing method (Kane et al., 2013b), and then training the frame-wise probability of creak along with normal vocoder parameters using SPSS. Thus, the statistical model learnt to predict the creaky voice from the contextual factors of the input text. Also, a specific f_0 estimation algorithm (Raitio et al., 2011c) is used and evaluated to enable the accurate estimation of the low-f_0 parts in the creak. In synthesis, the creaky excitation parts are rendered using creaky excitation instead of normal excitation and using the DSM vocoder (Drugman and Dutoit, 2012; Drugman et al., 2012b). Three voices are built to evaluate the described improvements in creaky synthesis: 1) the conventional system without special f_0 estimation or creaky excitation rendering, 2) an improved system with new f_0 estimation, 3) an improved system with new f_0 estimation and creaky voice rendering. The voices are evaluated

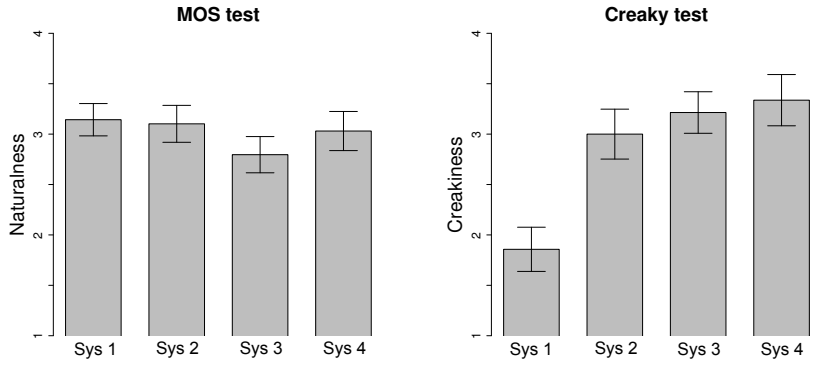


Figure 6.11. Subjective evaluation results of creaky transformation for the MOS (left) and the creaky preference test (right). The data is displayed as means and 95% confidence intervals.

in subjective listening tests, which indicate a clear improvement in naturalness when using the improved f_0 estimation and a preference for the system using the creaky excitation over the one without.

Finally, creaky voice transplantation methods are experimented with, meaning that a synthetic voice without creaky characteristics is augmented with a creaky voice capability. The transplantation strategy aims to transform two speech components, f_0 contours and creaky excitation, in order to enable creaky voice characteristics. The synthetic voice of the original non-creaky speaker and three different creaky-modified systems are evaluated in order to find out which strategy results in the perceptually most natural and creaky synthesis output: 1) the original speaker; 2) use of another speaker's creaky excitation in those parts where that speaker used a creaky voice; 3) in addition to 2), the use of a trained f_0 model from the creaky speaker; and 4) in addition to 2), the use of a data-driven f_0 transformation to model the f_0 contours for creaky voice. The evaluation results in terms of the MOS and overall creakiness are shown in Figure 6.11. The results show that all except system 3 are rated equally natural. System 3 is rated worse probably due to the inconsistent f_0 stream from another speaker. The creaky-modified systems 2–4 are all rated equally creaky and much more creaky than system 1, indicating successful creaky voice transplantation.

7. Conclusions

Speech synthesis is a difficult problem. Converting a given text to speech requires not only the successful synthesis of the linguistically important acoustic features, but also generating appropriate extralinguistic cues, such as speaking style and speaker characteristics. In combination, the linguistic and extralinguistic information finally define the ultimate meaning of an utterance. Thus, speech synthesis requires both the successful prediction of context (and meaning) from the input text and the modelling and reproduction of all perceptually relevant acoustic cues in speech. This thesis concentrates on the latter by giving special emphasis on the estimation and modelling of the voice source of speech.

This thesis presented two new glottal inverse filtering methods, that were shown to provide better glottal flow estimates than well-known reference methods. The latter GIF method was also shown to provide better speech quality than an existing GIF method, when used for analysis-synthesis of normal and shouted speech using GIF-based vocoding.

Several voice source modelling methods were also developed that utilise GIF-based analysis and synthesis of speech. All of them incorporate voice source analysis using GIF, and a detailed voice source parameterization thereof, which aims to enable more natural and expressive speech synthesis. A full vocoder utilising GIF was constructed, which incorporates several synthesis methods using the developed voice source modelling methods. The first method uses a glottal flow pulse extracted from natural speech using GIF, which is modified in synthesis to reconstruct the voiced excitation of speech. An important part of the modification is the spectral matching filter that allows the voice source to have a varying spectrum based on the context. Another developed method uses a library of windowed two-pitch-period glottal flow derivative pulses, which can be used for the unit selection type of voice source synthesis, where individual glot-

tal flow pulses are selected and concatenated based on their target and join costs. This allows the voice source to vary from one pitch period to another in order to reproduce more natural voice source characteristics. Finally, a third voice source modelling method was presented, which utilises a deep neural network to learn a mapping between extracted speech features and the time-domain glottal flow derivative waveform. The third method enables the automatic and robust generation of a voice source that is allowed to vary its spectrum and waveform shape from one pitch period to another. All of the methods were shown to yield high-quality synthetic speech based on the results of extensive subjective listening tests, and in many cases better quality was achieved when compared to well-known and widely used reference vocoding methods.

In addition to modelling the deterministic glottal flow pulse waveforms, a study was conducted to find the perceptual contribution and to develop synthesis methods for the aperiodic part of the voiced excitation. The study shows that a noise model is crucial for synthesis quality, and modeling of the time-varying spectrum of the noise is beneficial. The same study also shows that the phase characteristics of the excitation waveform have a significant perceptual effect in low-pitch speech.

Finally, the developed vocoding methods utilising GIF were used to synthesise various types of expressive speech. Firstly, the acoustic differences of breathy, normal, and Lombard speech were studied, and a synthesis scheme for generating speech with varying degrees of vocal effort was constructed. The resulting breathy, normal, and Lombard speech were evaluated in a series of extensive subjective listening tests in varying noise conditions. The results showed that the intelligibility of both natural and synthetic speech increases as the vocal effort is increased even if the loudness of the speech samples is equalised. The results also show that the synthetic speech with varying vocal effort was rated very similarly to natural speech with the corresponding variation in the vocal effort. This is backed up by the conducted suitability evaluations that show similar ratings for both synthetic and natural speech in different noise conditions. Secondly, the differences of normal and shouted speech were studied, and a synthesis scheme for shouted speech was developed. The study showed that spectral estimation methods that are not biased by the sparse harmonics of shouted speech is beneficial for synthesis quality. The experiments also showed that despite some quality degradation due to large differences in normal and shouted speech, the impression of shouting and

use of vocal effort was fairly well preserved in synthetic shout. Finally, a synthesis scheme for creaky voice was developed. The method consists of creaky voice detection, robust f_0 estimation, prediction of creaky voice from context, and rendering of the specific creaky excitation waveform in synthesis. Subjective listening tests showed that the synthetic creaky voices were rated more natural and more creaky compared to a conventional voice.

This thesis has shown several important aspects regarding statistical parametric speech synthesis. First, the modelling of the voice source signal is crucial for high-quality speech synthesis. Although the vocal tract shape and the resulting spectral filter is the most salient cue in speech, the voice source contributes considerably to the naturalness and expressivity of speech. An important aspect in the modelling of the voice source is the phase characteristics of the glottal flow waveform. This is especially important with low-pitch speech, where humans can perceive the phase characteristics of a repetitive signal, the voiced excitation of speech. In contrast, with speech of higher pitch, such as female speech, the phase correctness has less perceptual significance, but the modelling of the aperiodic component becomes the main factor for voice source quality.

For synthesising expressive speech, the modelling of the voice source waveform becomes even more important. However, although the modelling of the voice source signal has been shown to yield improvements in both segmental and suprasegmental quality, especially improving the segmental quality requires more than just the correct modelling of one pitch period, in particular, its phase characteristics. Since the repetitive patterns in speech are not perceived as individual units but in relation to each other, correct modelling of the modulation of the voice source from one pitch period to another is important. This is especially important for expressive speech, where the glottal vibrations show more irregularities than in modal speech. Such patterns are difficult to model using current techniques, but modelling approaches that meet such goals should be developed to improve the segmental quality of parametric speech synthesis. Similarly, an important aspect is the source-filter independence that is commonly assumed in methods utilising source-filter models. This assumption makes modelling easier, but also discards important correlations between the source and filter. In conventional HMM-based synthesis using a source-filter model, this correlation is lost, and as a result, synthetic speech with low segmental speech quality is produced. For example,

in conventional frame-based analysis of the source and filter, all the specific time-dependent details that occur during an excitation instant and during the time that follows before the next excitation are lost, although they have significant effects in the physiological speech production mechanism. These effects may significantly contribute to the segmental quality of speech. Regardless of whether this phenomenon is called source-filter dependence or source-filter interaction, the aim should be to link these two components more closely together. Finally, a key to successful speech synthesis is robustness. All the individual components in TTS synthesis, from data construction, recording, segmentation and labelling to parameter extraction, for example, should be performed as accurately as possible. If one step is not properly designed and performed, the whole chain will fail to produce the desired quality. Thus, speech synthesis needs careful planning, execution, and robust and well-functioning methods and algorithms.

In conclusion, the findings in this thesis support the use of data-driven voice source modelling methods in SPSS. Recently introduced automatic and robust tools for voice source analysis are most likely to propel the advancements in data-driven voice source modelling for SPSS. Combined with advanced machine learning techniques, this will probably have a significant impact on the quality of SPSS. Although the quality of SPSS has been increasing steadily, SPSS will probably not surpass the segmental quality of high-quality unit selection synthesis in the near future. However, the quality of SPSS will be acceptable for most products, and the strength of SPSS, such as the flexibility, will make it useful for various applications. In addition to improving segmental quality of synthetic speech, the main scientific and engineering challenges lie in the synthesis of expressive speech and specific voices, such as females and children, that exhibit a large variation in pitch, voice source spectrum, and excitation pattern. SPSS provides a feasible framework for synthesising various speaking styles and voices where unit selection methods are often impractical, but this also poses a challenge for voice source analysis and modelling. Steady progress has already been made in this area, but further research is required in order to get closer to the goal of speech synthesis, which is to generate synthetic speech that has the naturalness of real human speech and the ability to vary the type of expression and character depending on the context and requirements of the application.

References

Abercrombie, D. *Elements of General Phonetics*. Edinburgh University Press, 1967.

Agiomyrgiannakis, Y. and Rosec, O. ARX-LF-based source-filter methods for voice modification and transformation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3589–3592, 2009.

Airaksinen, M., Raitio, T., Story, B., and Alku, P. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):596–607, 2014.

Airaksinen, M. and Alku, P. Parameterization of the glottal source with the phase plane plot. In *Proceedings of Interspeech*, pages 96–100, 2014.

Airas, M. TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):49–64, 2008.

Airas, M., Alku, P., and Vainio, M. Laryngeal voice quality changes in expression of prominence in continuous speech. In *Proceedings of the 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pages 135–138, 2007.

Alku, P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2–3):109–118, 1992.

Alku, P., Strik, H., and Vilkman, E. Parabolic spectral parameter – A new method for quantification of the glottal flow. *Speech Communication*, 22:67–79, 1997.

Alku, P., Tiitinen, H., and Näätänen, R. A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology*, 110:1329–1333, 1999.

Alku, P., Bäckström, T., and Vilkman, E. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

Alku, P., Horacek, J., Airas, M., Griffond-Boitier, F., and Laukkanen, A.-M. Performance of glottal inverse filtering as tested by aeroelastic modelling of phonation and FE modelling of vocal tract. *Acta Acustica united with Acustica*, 92:717–724, 2006a.

Alku, P., Story, B., and Airas, M. Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production. *Folia Phoniatrica et Logopaedica*, 58(1): 102–113, 2006b.

Alku, P., Magi, C., Yrttiaho, S., Bäckström, T., and Story, B. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *Journal of the Acoustical Society of America*, 125:3289–3305, 2009.

Alku, P. Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650, 2011.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., and Story, B. H. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *Journal of the Acoustical Society of America*, 134(2):1295–1313, 2013.

Allen, J., Hunnicutt, M. S., Klatt, D. H., Armstrong, R. C., and Pisoni, D. B. *From Text to Speech: The MITalk System*. Cambridge University Press, New York, NY, USA, 1987.

American National Standards Institute. Methods for calculation of the speech intelligibility index. ANSI S3.5-1997, 1997.

Ananthapadmanabha, T. and Yegnanarayana, B. Epoch extraction of voiced speech. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 23(6):562–570, 1975.

Ananthapadmanabha, T. and Yegnanarayana, B. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 27(4):309–319, 1979.

Ananthapadmanabha, T. V. Acoustic analysis of voice source dynamics. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 2–3:1–24, 1984.

Ananthapadmanabha, T. and Fant, G. Calculation of true glottal flow and its components. *Speech Communication*, 1(3–4):167–184, 1982.

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. A compact model for speaker-adaptive training. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 1137–1140, 1996.

Anthony, J. and Lawrence, W. A resonance analogue speech synthesizer. In *Proceedings of the 4th International Congress on Acoustics*, pages 1–2, 1962.

Astrinaki, M., Moinet, A., Yamagishi, J., Richmond, K., Ling, Z.-H., King, S., and Dutoit, T. Mage – HMM-based speech synthesis reactively controlled by the articulators. In *Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW8)*, pages 207–211, 2013.

Atal, B. S. and Hanauer, S. L. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50 (2B):637–655, 1971.

Auvinen, H., Raitio, T., Siltanen, S., and Alku, P. Utilizing Markov chain Monte Carlo (MCMC) method for improved glottal inverse filtering. In *Proceedings of Interspeech*, pages 1640–1643, 2012.

Auvinen, H., Raitio, T., Airaksinen, M., Siltanen, S., Story, B. H., and Alku, P. Automatic glottal inverse filtering with the Markov chain Monte Carlo method. *Computer Speech and Language*, 28(5):1139–1155, 2014.

Aylett, M. and Yamagishi, J. Combining statistical parametric speech synthe-

sis and unit-selection for automatic voice cloning. In *Proceedings of LangTech*, 2008.

Babacan, O., Drugman, T., Raitio, T., Erro, D., and Dutoit, T. Parametric representation for singing voice synthesis: A comparative evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2583–2587, 2014.

Bäckström, T., Alku, P., and Vilkman, E. Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range. *IEEE Transactions on Speech and Audio Processing*, 10(2): 186–192, 2002.

Bahl, L., Brown, P., De Souza, P. V., and Mercer, R. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 11, pages 49–52, 1986.

Bahl, L. R., Jelinek, F., and Mercer, R. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.

Baum, L. E. and Eagon, J. A. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, 1967.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S. The original ToBI system and the evolution of the ToBI framework. In Jun, S.-A., editor, *Prosodic Typology – The Phonology of Intonation and Phrasing*, chapter 2, pages 9–54. Oxford University Press, 2005.

Beerends, J., Larsen, E., Lyer, N., and van Vugt, J. Measurement of speech intelligibility based on the PESQ approach. In *Proceedings of the Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN)*, 2004.

Beerends, J., van Wijngaarden, S., and van Buuren, R. Extension of ITU-T recommendation P.862 PESQ towards measuring speech intelligibility with vocoders. In *Proceedings of NATO Research and Technology Organisation (RTO), Human Factors and Medicine Panel (MP-HFM-123): New Directions for Improving Audio Effectiveness*, pages 10-1–10-6, 2005.

Bellegarda, J. R. and Nahamoo, D. Tied mixture continuous parameter modeling for speech recognition. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 38(12):2033–2045, 1990.

Beller, G., Obin, N., and Rodet, X. Articulation degree as a prosodic dimension of expressive speech. In *Proceedings of the 4th International Conference on Speech Prosody*, 2008.

Benoît, C., Grice, M., and Hazan, V. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18(4):381–392, 1996.

Bilsen, F. A. On the influence of the number and phase of harmonics on the perceptibility of the pitch of complex signals. *Acustica*, 28:60–65, 1973.

Black, A. W. Unit selection and emotional speech. In *Proceedings of Eurospeech*, pages 1649–1652, 2003.

Black, A. W. and Campbell, N. Optimising selection of units from speech database for concatenative synthesis. In *Proceedings of Eurospeech*, pages 581–584, 1995.

Black, A. W. and Taylor, P. CHATR: A generic speech synthesis system. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, volume 2, pages 983–986, 1994.

Black, A. W. and Taylor, P. Festival speech synthesis system: System documentation (1.1.1). Techinal report HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, 1997a.

Black, A. W. and Tokuda, K. Blizzard challange – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proceedings of Interspeech*, pages 77–80, 2005.

Black, A. W., Zen, H., and Tokuda, K. Statistical parametric speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 1229–1232, 2007.

Black, A. W., Bunnell, H. T., Dou, Y., Kumar Muthukumar, P., Metze, F., Perry, D., Polzehl, T., Prahallad, K., Steidl, S., and Vaughn, C. Articulatory features for expressive speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4005–4008, 2012.

Black, A. W. and Taylor, P. Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings of Eurospeech*, pages 601–604, 1997b.

Black, A. W., Lenzo, K., and Pagel, V. Issues in building general letter to sound rules. In *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pages 77–80, 1998.

Black, A. W., Taylor, P., Caley, R., Clark, R., Richmond, K., King, S., Strom, V., and Zen, H. The Festival Speech Synthesis System Version 1.4.2. Software, 2001. URL `http://www.cstr.ed.ac.uk/projects/festival/`.

Blizzard Challenge. The Blizzard Challenge, 2014. URL `http://www.synsig.org/index.php/Blizzard_Challenge`.

Blomgren, M., Chen, Y., Ng, M. L., and Gilbert, H. R. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *Journal of the Acoustical Society of America*, 103(5):2649–2658, 1998.

Bouzid, A. and Ellouze, N. Glottal opening instant detection from speech signal. In *Proceedings of the 14th European Signal Processing Conference (EUSIPCO)*, pages 729–732, 2004.

Bozkurt, B. and Dutoit, T. Mixed-phase speech modeling and formant estimation using differential phase spectrums. In *Voice Quality: Functions, Analysis and Synthesis, ISCA Tutorial and Research Workshop (VOQUAL '03)*, pages 21–24, 2003.

Bozkurt, B., Doval, B., d'Alessandro, C., and Dutoit, T. Zeros of z-transform representation with application to source-filter separation in speech. *IEEE Signal Processing Letters*, 12(4):344–347, 2005.

Bozkurt, B., Couvreur, L., and Dutoit, T. Chirp group delay analysis of speech signals. *Speech Communication*, 49:159–176, 2007.

Breen, A. P. and Jackson, P. Non-uniform unit selection and the similar-

ity metric within BT's Laureate TTS system. In *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pages 201–206, 1998.

Brokx, J. P. L. and Nooteboom, S. G. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10:23–36, 1981.

Brookes, D. M. and Loke, H. P. Modelling energy flow in the vocal tract with applications to glottal closure and opening detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 213–216, 1999.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. Class-based N-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

Cabral, J., Renalds, S., Richmond, K., and Yamagishi, J. Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, pages 113–118, 2007.

Cabral, J., Renalds, S., Richmond, K., and Yamagishi, J. Glottal spectral separation for parametric speech synthesis. In *Proceedings of Interspeech*, pages 1829–1832, 2008.

Cabral, J. P., Kane, J., Gobl, C., and Carson-Berndsen, J. Evaluation of glottal epoch detection algorithms on different voice types. In *Proceedings of Interspeech*, pages 1989–1992, 2011a.

Cabral, J. P., Renals, S., Yamagishi, J., and Richmond, K. HMM-based speech synthesiser using the LF-model of the glottal source. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4704–4707, 2011b.

Cabral, J. P., Richmond, K., Yamagishi, J., and Renals, S. Glottal spectral separation for speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):195–208, 2014.

Campbell, N. and Black, A. W. Prosody and the selection of source units for concatenative synthesis. In van Santen, J., Sproat, R., Olive, J., and Hirschberg, J., editors, *Progress in speech synthesis*, pages 279–282. Springer-Verlag, New York, 1996.

Campbell, N. and Mokhtari, P. Voice quality: The 4th prosodic dimension. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2417–2420, 2003.

Campbell, N. Evaluation of speech synthesis. From reading machines to talking machines. In Dybkjær, L., Hemsen, H., and Minker, W., editors, *Evaluation of Text and Speech Systems*, Text, Speech and Language Technology, pages 29–64. Springer, The Netherlands, 2007.

Campillo, F., Méndez, F., Arza, M., Docío, L., Bonafonte, A., Navas, E., and Sainz, I. Albayzín 2010: A Spanish text to speech evaluation. In *Proceedings of Interspeech*, pages 2161–2164, 2011.

Carlson, R., Granström, B., and Klatt, D. Vowel perception: The relative perceptual salience of selected acoustic manipulations. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 20(3–4):73–83, 1979.

Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I., and Lin, Q.-G. Voice source rules for text-to-speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*

(ICASSP), volume 1, pages 223–226, 1989.

Carlson, R., Granström, B., and Karlsson, I. Experiments with voice modelling in speech synthesis. *Speech Communication*, 10:481–489, 1991.

Catford, J. C. *Fundamental Problems in Phonetics*. Edinburgh University Press, Edinburgh, 1977.

Cerňak, M. and Rusko, M. An evaluation of a synthetic speech using the PESQ measure. In *Proceedings of Forum Acusticum*, 2005.

Cerňak, M., Milan, R., and Marian, T. Diagnostic evaluation of synthetic speech using speech recognition. In *Proceedings of the 16th International Congress on Sound and Vibration*, 2009.

Chang, Y.-Y. Evaluation of TTS systems in intelligibility and comprehension tasks. In *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 64–78, 2011.

Charpentier, F. and Stella, M. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 11, pages 2015–2018, 1986.

Chen, J.-H. and Gersho, A. Adaptive postfiltering for quality enhancement of coded speech. *IEEE Transactions on Speech and Audio Processing*, 3 (1):59–71, 1995.

Chen, L.-H., Yang, C.-Y., Ling, Z.-H., Jiang, Y., Dai, L.-R., Hu, Y., and Wang, R.-H. The USTC system for Blizzard Challenge 2011. In *Proceedings of Blizzard Challenge Workshop*, 2011. URL `http://festvox.org/blizzard`.

Chen, L.-H., Ling, Z.-H., Song, Y. J. Y., Xia, X.-J., Zu, Y.-Q., Yan, R.-Q., and Dai, L.-R. The USTC system for Blizzard Challenge 2013. In *Proceedings of Blizzard Challenge Workshop*, 2013. URL `http://festvox.org/blizzard`.

Chen, L.-H., Raitio, T., Valentini-Botinhao, C., Yamagishi, J., and Ling, Z.-H. DNN-based stochastic postfilter for HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 1954–1958, 2014.

Cheng, Y. and O'Shaughnessy, D. Automatic and reliable estimation of glottal closure instant and period. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 37(12):1805–1815, 1989.

Childers, D. Glottal source modeling for voice conversion. *Speech Communication*, 16(2):127–138, 1995.

Childers, D. and Ahn, C. Modeling the glottal volume-velocity waveform for three voice types. *Journal of the Acoustical Society of America*, 72(1):505–519, 1995.

Childers, D. and Hu, H. Speech synthesis by glottal excited linear prediction. *Journal of the Acoustical Society of America*, 96(4):2026–2036, 1994.

Childers, D. G. and Lee, C. K. Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90(5):2394–2410, 1991.

Christiansen, C., Pedersen, M. S., and Dau, T. Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communication*, 52 (7–8):678–692, 2010.

Clark, R. A. J. and Dusterhoff, K. E. Objective methods for evaluating synthetic intonation. In *Proceedings of Eurospeech*, pages 1623–1626, 1999.

Cole, R. A. and Scott, B. Toward a theory of speech perception. *Psychological*

Review, 81(4):348–374, 1974.

Cooke, M., Mayo, C., and Valentini-Botinhao, C. Intelligibility-enhancing speech modifications: The Hurricane Challenge. In *Proceedings of Interspeech*, pages 3552–3556, 2013.

Cooke, M. A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 119(3):1562–1573, 2006.

Dau, T., Püschel, D., and Kohlrausch, A. A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *Journal of the Acoustical Society of America*, 99(6):3615–3622, 1996.

Davis, S. B. Acoustic characteristics of normal and pathological voices. *Speech and language: Advances in basic research and practice*, 1:271–335, 1979.

de Boer, E. A note on phase distortion and hearing. *Acustica*, 11:182–184, 1961.

de Veth, J., Cranen, B., Strik, H., and Boves, L. Extraction of control parameters for the voice source in a text-to-speech system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 301–304, 1990.

Degottex, G., Roebel, A., and Rodet, X. Phase minimization for glottal model estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1080–1090, 2011.

Degottex, G. and Erro, D. A uniform phase representation for the harmonic model in speech synthesis applications. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):38, 2014.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., and Acero, A. Recent advances in deep learning for speech research at Microsoft. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8604–8608, 2013.

Ding, W., Campbell, N., Higuchi, N., and Kasuya, H. Fast and robust joint estimation of vocal tract and voice source parameters. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1291–1294, 1997.

Ding, W. and Campbell, N. Determining polarity of speech signals based on gradient of spurious glottal waveforms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 857–860, 1998.

Do, C.-T., Evrard, M., Leman, A., d'Alessandro, C., Rilliard, A., and Crebouw, J.-L. Objective evaluation of HMM-based speech synthesis system using Kullback–Leibler divergence. In *Proceedings of Interspeech*, pages 2952–2956, 2014.

Dromey, C., Stathopoulos, E., and Sapienza, C. Glottal airflow and electroglottographic measures of vocal function at multiple intensities. *Journal of Voice*, 6(1):44–54, 1992.

Drugman, T. *Advances in glottal analysis and its applications*. PhD thesis, University of Mons, Belgium, 2011.

Drugman, T. Residual excitation skewness for automatic speech polarity detec-

tion. *IEEE Signal Processing Letters*, 20(4):387–390, 2013.

Drugman, T. and Dutoit, T. Glottal closure and opening instant detection from speech signals. In *Proceedings of Interspeech*, pages 2891–2894, 2009.

Drugman, T. and Dutoit, T. The deterministic plus stochastic model of the residual signal and its applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):968–981, 2012.

Drugman, T. and Dutoit, T. Detecting speech polarity with high-order statistics. *Cognitive Computation*, 5(4):442–447, 2013.

Drugman, T. and Stylianou, Y. Maximum voiced frequency estimation: Exploiting amplitude and phase spectra. *IEEE Signal Processing Letters*, 21 (10):1230–1234, 2014.

Drugman, T., Bozkurt, B., and Dutoit, T. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proceedings of Interspeech*, pages 116–119, 2009a.

Drugman, T., Wilfart, G., and Dutoit, T. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proceedings of Interspeech*, pages 1779–1782, 2009b.

Drugman, T., Wilfart, G., Moinet, A., and Dutoit, T. Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3793–3796, 2009c.

Drugman, T., Bozkurt, B., and Dutoit, T. A comparative study of glottal source estimation techniques. *Computer Speech and Language*, 26(1):20–34, 2012a.

Drugman, T., Kane, J., and Gobl, C. Modeling the creaky excitation for parametric speech synthesis. In *Proceedings of Interspeech*, 2012b.

Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3): 994–1006, 2012c.

Drugman, T., Kane, J., Raitio, T., and Gobl, C. Prediction of creaky voice from contextual factors. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7967–7971, 2013.

Drugman, T. and Dutoit, T. Speech polarity determination: A comparative evaluation. *Neurocomputing*, 132(0):121–125, 2014.

Drugman, T. and Raitio, T. Excitation modeling for HMM-based speech synthesis: Breaking down the impact of periodic and aperiodic components. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 260–264, 2014.

Drugman, T., Bozkurt, B., and Dutoit, T. Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. *Speech Communication*, 53(6):855–866, 2011.

Drugman, T., Alku, P., Alwan, A., and Yegnanarayana, B. Glottal source processing: From analysis to applications. *Computer Speech and Language*, 28 (5):1117–1138, 2014.

Dudley, H. The vocoder. *Bell Labs Record*, 17:122–126, 1939.

Duffy, S. A. and Pisoni, D. B. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*,

35(4):351–389, 1992.

Dutoit, T. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, 1997.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 1393–1396, 1996.

Edwards, J. A. and Angus, J. A. S. Using phase-plane plots to assess glottal inverse filtering. *Electronics Letters*, 32(3):192–193, 1996.

El-Jaroudi, A. and Makhoul, J. Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39(2):411–423, 1991.

Elliott, J. Comparing the acoustic properties of normal and shouted speech: A study in forensic phonetics. In *Proceedings of the 8th International Conference on Speech Science and Technology (SST-2000)*, pages 154–159, 2000.

Erro, D., Sainz, I., Navas, E., and Hernaez, I. Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):184–194, 2014.

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

Falk, T. H., Möller, S., Karaiskos, V., and King, S. Improving instrumental quality prediction performance for the Blizzard Challenge. In *Proceedings of Blizzard Challenge Workshop*, 2008. URL `http://festvox.org/blizzard`.

Fan, Y. Q. Y., Hu, W., and Soong, F. K. On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3857–3861, 2014a.

Fan, Y., Qian, Y., Xie, F., and Soong, F. K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Proceedings of Interspeech*, pages 1964–1968, 2014b.

Fant, G. Speech communication research. *Royal Swedish Academy of Engineering Sciences (IVA)*, 24(8):331–337, 1953.

Fant, G. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960. 2nd edition 1970.

Fant, G. Some problems in voice source analysis. *Speech Communication*, 13: 7–22, 1993.

Fant, G. The LF-model revisited. Transformations and frequency domain analysis. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 36(2–3):119–156, 1995.

Fant, G. and Lin, Q. Glottal source – Vocal tract acoustic interaction. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 28(1):13–27, 1987.

Fant, G. and Martony, J. Speech synthesis. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 2:18–24, 1962.

Fant, G., Liljencrants, J., and Lin, Q. A four-parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 26(4):1–13, 1985a.

Fant, G., Lin, Q., and Gobl, C. Notes on glottal flow interaction. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 2–3:21–45, 1985b.

Fant, G., Kruckenberg, A., Liljencrants, J., and Båvegård, M. Voice source parameters in continuous speech. Transformation of LF-parameters. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 1451–1454, 1994.

Fant, G. The voice source in connected speech. *Speech Communication*, 22(2–3): 125–139, 1997.

Fernandez, R., Rendel, A., Ramabhadran, B., and Hoory, R. F0 contour prediction with a deep belief network-Gaussian process hybrid model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6885–6889, 2013.

Flanagan, J. *Speech Analysis, Synthesis and Perception*. Springer-Verlag, Berlin/Heidelberg/New York, 2nd edition, 1972a.

Flanagan, J. L. Voices of men and machines. *Journal of the Acoustical Society of America*, 51(5A):1375–1387, 1972b.

Flanagan, J. L. Computers that talk and listen: Man-machine communication by voice. *Proceedings of the IEEE*, 64(4):405–415, 1976.

Fletcher, H. Loudness, masking and their relation to the hearing process and the problem of noise measurement. *Journal of the Acoustical Society of America*, 9:275–293, 1938a.

Fletcher, H. The mechanism of hearing as revealed through experiments on the masking effect of thermal noise. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 24:265–274, 1938b.

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

Forney, G. D., Jr. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

Franz, A. and Brants, T. All our N-gram are belong to you. Google Research Blog, 2006. URL `http://googleresearch.blogspot.fi/2006/08/all-our-n-gram-are-belong-to-you.html`.

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

Fries, G. Hybrid time- and frequency-domain speech synthesis with extended glottal source generation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 581–584, 1994.

Fröhlich, M., Michaelis, D., and Strube, H. SIM – Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *Journal of the Acoustical Society of America*, 110(1):479–488, 2001.

Frøkjær-Jensen, B. and Prytz, S. Registration of voice quality. *Brüel&Kjær Technical Review*, 3:3–17, 1973.

Fu, Q. and Murphy, P. Adaptive inverse filtering for high accuracy estimation

of the glottal source. In *Proceedings of the ISCA Tutorial and Research Workshop, Non-Linear Speech Processing, Paper 018*, 2003.

Fu, Q. and Murphy, P. Robust glottal source estimation based on joint source-filter model optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):492–501, 2006.

Fujinaga, K., Nakai, M., Shimodaira, H., and Sagayama, S. Multiple-regression hidden Markov model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 513–516, 2001.

Fujisaki, H. and Ljungqvist, M. Proposal and evaluation of models for the glottal source waveform. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 11, pages 1605–1608, 1986.

Fujisaki, H. and Ljungqvist, M. Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 637–640, 1987.

Fujisaki, H., Hirose, K., Halle, P., and Lei, H. A generative model for the prosody of connected speech in Japanese. *Annual Report of Engineering Research Institute*, 30:75–80, 1971.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. An adaptive algorithm for mel-cepstral analysis of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 137–140, 1992.

Furui, S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 34(1):52–59, 1986.

Gamerman, D. *Markov Chain Monte Carlo – Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, Boca Raton/London/New York/ Washington, D.C., 1997.

Gauvain, J. and Lee, C.-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, Boca Raton/London/New York/ Washington, D.C., 1996.

Gobl, C. A preliminary study of acoustic voice quality correlates. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 4:9–12, 1989.

Gobl, C. and Ní Chasaide, A. Acoustic characteristics of voice quality. *Speech Communication*, 11:481–490, 1992.

Gobl, C. and Ní Chasaide, A. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1-2):189–212, 2003.

Gómez-Vilda, P., Fernández-Baillo, R., Rodellar-Biarge, V., Lluis, V. N., Álvarez-Marquina, A., Mazaira-Fernández, L. M., Martínez-Olalla, R., and Godino-Llorente, J. I. Glottal source biometrical signature for voice pathology detection. *Speech Communication*, 51(9):759–781, 2009.

Gordon, M. and Ladefoged, P. Phonation types: A cross-linguistic review. *Journal*

of Phonetics, 29(4):383–406, 2001.

Gray, H. *Anatomy of the Human Body*. Lea & Febiger, Philadelphia, 1918.

Gray, J., A. and Markel, J. Distance measures for speech processing. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 24(5):380–391, 1976.

Gray, R., Buzo, A., Gray, J., A., and Matsuyama, Y. Distortion measures for speech processing. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 28(4):367–376, 1980.

Griffin, D. W. and Lim, J. S. Multiband excitation vocoder. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 36(8):1223–1235, 1988.

Gudnason, J., Thomas, M., Naylor, P., and Ellis, D. Voice source waveform analysis and synthesis using principal component analysis and Gaussian mixture modelling. In *Proceedings of Interspeech*, pages 108–111, 2009.

Gudnason, J., Thomas, M. R. P., Ellis, D. P. W., and Naylor, P. A. Data-driven voice source waveform analysis and synthesis. *Speech Communication*, 54(2):199–211, 2012.

Hacki, T. Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. *Folia Phoniatrica*, pages 43–48, 1989.

Handley, Z. Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication*, 51(10):906–919, 2009.

Hanson, H. Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, 101(1):466–481, 1997.

Hanson, H. and Chuang, E. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America*, 106(1):1064–1077, 1999.

Hanson, H. M. *Glottal characteristics of female speakers*. PhD thesis, Harvard University, United States, 1995.

Harma, A. and Laine, U. K. A comparison of warped and conventional linear predictive coding. *IEEE Transactions on Speech and Audio Processing*, 9(5):579–588, 2001.

Hastings, W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

Hedelin, P. A glottal LPC-vocoder. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 9, pages 21–24, 1984.

Hedelin, P. and Huber, D. Pitch period determination of aperiodic speech signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 361–364, 1990.

Henter, G. E., Merritt, T., Shannon, M., Mayo, C., and King, S. Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech. In *Proceedings of Interspeech*, pages 1504–1508, 2014.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5, Pt. 1):3099–3111, 1995.

Hillenbrand, J. and Gayvert, R. T. Identification of steady-state vowels synthesized from the Peterson and Barney measurements. *Journal of the Acoustical Society of America*, 94(2):668–674, 1993.

Hillman, R., Holmberg, E., Perkell, J., Walsh, M., and Vaughan, C. Objective assessment of vocal hyperfunction: An experimental framework and initial results. *Journal of Speech, Language, and Hearing Research*, 32: 373–392, 1989.

Hinterleitner, F., Möller, S., Falk, T. H., and Polzehl, T. Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from Blizzard Challenges 2008 and 2009. In *Proceedings of Blizzard Challenge Workshop*, 2010. URL `http://festvox.org/blizzard`.

Hinterleitner, F., Möller, S., Norrenbrock, C. R., and Heute, U. Perceptual quality dimensions of text-to-speech systems. In *Proceedings of Interspeech*, pages 2177–2180, 2011.

Hinterleitner, F., Norrenbrock, C. R., and Möller, S. Is intelligibility still the main problem? A review of perceptual quality dimensions of synthetic speech. In *Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW8)*, pages 147–151, 2013.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Hirai, T. and Tenpaku, S. Using 5 ms segments in concatenative speech synthesis. In *Proceedings of the 5th ISCA Workshop on Speech Synthesis (SSW5)*, 2004.

Hirst, D., Rilliard, A., and Aubergé, V. Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. In *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, 1998.

Holmberg, E., Hillman, R., and Perkell, J. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *Journal of the Acoustical Society of America*, 84(2):511–529, 1988.

Holmes, J. The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE Transactions on Audio and Electroacoustics*, 21(3):298–305, 1973.

Hon, H.-W., Acero, A., Huang, X.-D., Liu, J.-S., and Plumpe, M. Automatic generation of synthesis units for trainable text-to-speech systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 293–296, 1998.

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

Howell, P. and Williams, M. The contribution of the excitatory source to the perception of neutral vowels in stuttered speech. *Journal of the Acoustical Society of America*, 84(1):80–89, 1988.

Howell, P. and Williams, M. Acoustic analysis and perception of vowels in children's and teenagers' stuttered speech. *Journal of the Acoustical Society of America*, 91(3):1697–1706, 1992.

HTS. HMM-based Speech Synthesis System (HTS). Software, 2014. URL `http://hts.sp.nitech.ac.jp`.

Hu, Q., Stylianou, Y., Maia, R., Richmond, K., Yamagishi, J., and Latorre, J. An investigation of the application of dynamic sinusoidal models to statistical parametric speech synthesis. In *Proceedings of Interspeech*, pages 780–784, 2014.

Hu, Y. and Loizou, P. C. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

Huang, D.-Y. Prediction of perceived sound quality of synthetic speech. In *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2011.

Huang, X., Acero, A., and Hon, H.-W. *Spoken Language Processing. A guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, New Jersey, 2001.

Huang, X.-D., Acero, A., Adcock, J., Hon, H.-W., Goldsmith, J., and Liu, J.-S. Whistler: A trainable text-to-speech system. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 2387–2390, 1996.

Huber, S., Roebel, A., and Degottex, G. Glottal source shape parameter estimation using phase minimization variants. In *Proceedings of Interspeech*, 2012.

Hunt, A. and Black, A. W. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 373–376, 1996.

IEEE Subcommittee on Subjective Measurements. IEEE recommended practices for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17:227–246, 1969.

Imai, S. Cepstral analysis synthesis on the mel frequency scale. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 8, pages 93–96, 1983.

International Phonetic Association. IPA Chart, 2005. URL `http://www.langsci.ucl.ac.uk/ipa/ipachart.html`. Available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2005 International Phonetic Association.

International Telecommunication Union. Methods for subjective determination of transmission quality. Recommendation ITU-T P.800 (08/96), 1996.

International Telecommunication Union. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Recommendation ITU-T P.862 (02/2001), 1997.

Iseli, M., Shue, Y.-L., and Alwan, A. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *Journal of the Acoustical Society of America*, 121(4):2283–2295, 2007.

Ishi, C. T., Ishiguro, H., and Hagita, N. Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 3:1–12, 2010.

Itakura, F. and Saito, S. Analysis synthesis telephony based upon the maximum likelihood method. In *Proceedings of the 6th International Congress on Acoustics*, volume C-5-5, pages 17–20, 1968.

Iwarsson, J., Thomasson, M., and Sundberg, J. Effects of lung volume on glottal voice source. *Journal of Voice*, 12(4):424–433, 1998.

Jiang, Y., Ling, Z.-H., Lei, M., Wang, C.-C., Lu, H., Hu, Y., Dai, L.-R., and Wang, R.-H. The USTC system for Blizzard Challenge 2010. In *Proceedings of Blizzard Challenge Workshop*, 2010. URL `http://festvox.org/blizzard`.

Joris, P. X., Carney, L. H., Smith, P. H., and Yin, T. C. Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency. *Journal of Neurophysiology*, 71 (3):1022–1036, 1994.

Juang, B.-H. and Rabiner, L. Mixture autoregressive hidden Markov models for speech signals. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 33(6):1404–1413, 1985.

Junqua, J.-C. The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 93(1): 510–524, 1993.

Kadambe, S. and Bourdreaux-Batels, G. Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory*, 32(2):917–924, 1992.

Kaipio, J. and Somersalo, E. *Statistical and Computational Inverse Problems*. Springer-Verlag, New York, USA, 2005.

Kane, J. and Gobl, C. Evaluation of glottal closure instant detection in a range of voice qualities. *Speech Communication*, 55(2):295–314, 2013.

Kane, J., Scherer, S., Morency, L.-P., and Gobl, C. A comparative study of glottal open quotient estimation techniques. In *Proceedings of Interspeech*, pages 1658–1662, 2013a.

Kane, J. *Tools for analysing the voice: Developments in glottal source and voice quality analysis*. PhD thesis, Trinity College Dublin, Ireland, 2012.

Kane, J., Drugman, T., and Gobl, C. Improved automatic detection of creak. *Computer Speech and Language*, 27(4):1028–1047, 2013b.

Kang, S., Qian, X., and Meng, H. Multi-distribution deep belief network for speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8012–8016, 2013.

Karhila, R., Remes, U., and Kurimo, M. Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):285–295, 2014.

Karjalainen, M., Laine, U., Toivonen, R., Haymond, K., Folmar, R., and Wood, J. Aids for the handicapped based on "Synte 2" speech synthesizer. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 851–854, 1980.

Karjalainen, M., Altosaar, T., and Vainio, M. Speech synthesis using warped linear prediction and neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 877–880, 1998.

Karlsson, I. Female voices in speech synthesis. *Journal of Phonetics*, 19:111–120, 1991.

Karlsson, I. Modelling voice variations in female speech synthesis. *Speech Communication*, 11(4–5):491–495, 1992.

Kasuya, H., Maekawa, K., and Kiritani, S. Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics. In *Proceedings of the 14th International Congress of Phonetic Sciences*, volume 3, pages 2505–2512, 1999.

Kates, J. and Arehart, K. Coherence and the speech intelligibility index. *Journal of the Acoustical Society of America*, 115(5):2604–2604, 2004.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3–4):187–207, 1999.

Kawahara, H., Estill, J., and Fujimura, O. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proceedings of the 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

Kawai, H., Toda, T., Ni, J., Tsuzaki, M., and Tokuda, K. XIMERA: A new TTS from ATR based on corpus-based technologies. In *Proceedings of the 5th ISCA Workshop on Speech Synthesis (SSW5)*, 2004.

Kelly, J. L. and Lochbaum, C. Speech synthesis. In *Proceedings of the Speech Communications Seminar. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm*, 1962.

Kent, R. D. and Read, C. *The Acoustic Analysis of Speech*. Singular Publishing Group, Inc., San Diego/London, 1992.

Kim, S.-J. and Hahn, M. Two-band excitation for HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, E90-D, 2007.

King, S. Measuring a decade of progress in text-to-speech. *Loquens*, 1(1), 2014.

King, S. and Karaiskos, V. The Blizzard Challenge 2012. In *Proceedings of Blizzard Challenge Workshop*, 2012. URL `http://festvox.org/blizzard`.

Kinnunen, T., Karpov, E., and Franti, P. Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):277–288, 2006.

Kinnunen, T. and Li, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40, 2010.

Klatt, D. H. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3):971–995, 1980.

Klatt, D. H. The Klattalk text-to-speech system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1589–1592, 1982.

Klatt, D. H. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3):737–793, 1987.

Klatt, D. H. and Klatt, L. C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

Klatt, D. H. Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. *Journal of the Acoustical Society of America*, 53(1):8–16, 1973.

Kominek, J. and Black, A. W. The Blizzard Challenge 2006 CMU entry intro-

ducing hybrid trajectory-selection synthesis. In *Proceedings of Blizzard Challenge Workshop*, 2006. URL `http://festvox.org/blizzard`.

Kounoudes, A., Naylor, P., and Brookes, M. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 349–352, 2002.

Kreiman, J., Gerratt, B., and Antonanzas-Barroso, N. Measures of the glottal source spectrum. *Journal of Speech, Language, and Hearing Research*, 50:595–610, 2007.

Krishnamurthy, A. and Childers, D. Two-channel speech analysis. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 34:730–743, 1986.

Krstulović, S., Hunecke, A., and Schroder, M. An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. In *Proceedings of Interspeech*, pages 1897–1900, 2007.

Kryter, K. D. Methods for the calculation and use of the articulation index. *Journal of the Acoustical Society of America*, 34(11):1689–1697, 1962.

Kubichek, R. F. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, volume 1, pages 125–128, 1993.

Kuhn, R., Junqua, J.-C., Nguyen, P., and Niedzielski, N. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000.

Ladefoged, P. *Preliminaries to Linguistic Phonetics*. University of Chicago, Chicago, 1971.

Laine, U. K., Karjalainen, M., and Altosaar, T. Warped linear prediction (WLP) in speech and audio processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 349–352, 1994.

Laitinen, M.-V., Disch, S., and Pulkki, V. Sensitivity of human hearing to changes in phase spectrum. *Journal of the Audio Engineering Society*, 61(11): 860–877, 2013.

Lanchantin, P., Degottex, G., and Rodet, X. A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4630–4633, 2010.

Larar, J., Alsaka, Y., and Childers, D. Variability in closed phase analysis of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 10, pages 1089–1092, 1985.

Laver, J. *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980.

Laver, J. and Trudgill, P. Phonetic and linguistic markers in speech. In Scherer, K. R. and Giles, H., editors, *Social markers in speech*, pages 1–32. Cambridge University Press, 1979.

Lawrence, W. The synthesis of speech from signals which have a low information rate. In Jackson, W., editor, *Communication Theory*, pages 460–469. Butterworths, London, England, 1953.

Leggetter, C. J. and Woodland, P. C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

Lei, M., Wu, Y.-J., Ling, Z.-H., and Dai, L.-R. Investigation of prosodic F0 layers in hierarchical F0 modeling for HMM-based speech synthesis. In *Proceedings of the IEEE 10th International Conference on Signal Processing (ICSP)*, pages 613–616, 2010.

Léon, P. and Martin, P. Machines and measurements. In Bolinger, D., editor, *Intonation*, pages 30–47. Penguin, Harmondsworth, UK, 1972.

Levinson, S. E. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1(1):29–45, 1986.

Liao, H. Speaker adaptation of context dependent deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7947–7951, 2013.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932.

Lin, Q. Nonlinear interaction in voice production. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 28(1):1–12, 1987.

Lin, Q. *Speech production theory and articulatory speech synthesis*. PhD thesis, Royal Institute of Technology, Stockholm, 1990.

Lindblom, B. Economy of speech gestures. In MacNeilage, P. F., editor, *The Production of Speech*, pages 217–245. Springer New York, 1983.

Ling, Z.-H. and Wang, R.-H. HMM-based unit selection using frame sized speech segments. In *Proceedings of Interspeech*, pages 2034–2037, 2006.

Ling, Z.-H. and Wang, R.-H. HMM-based hierarchical unit selection combining Kullback–Leibler divergence with likelihood criterion. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1245–1248, 2007.

Ling, Z.-H., Wu, Y., Wang, Y.-P., Qin, L., and Wang, R.-H. USTC system for Blizzard Challenge 2006: An improved HMM-based speech synthesis method. In *Proceedings of Blizzard Challenge Workshop*, 2006. URL `http://festvox.org/blizzard`.

Ling, Z.-H., Qin, L., Lu, H., Gao, Y., Dai, L.-R., Wang, R.-H., Jian, Y., Zhao, Z.-W., Yang, J.-H., Chen, J., and Hu, G.-P. The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007. In *Proceedings of Blizzard Challenge Workshop*, 2007. URL `http://festvox.org/blizzard`.

Ling, Z.-H., Zhang, W., and Wang, R.-H. Cross-stream dependency modeling for HMM-based speech synthesis. In *Proceedings of the 6th International Symposium on Chinese Spoken Language Processing*, pages 5–8, 2008.

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1171–1185, 2009.

Ling, Z.-H., Deng, L., and Yu, D. Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2129–2139, 2013.

Linggard, R. *Electronic Synthesis of Speech*. Cambridge University Press, Cambridge, 1985.

Lisker, L. and Abramson, A. S. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20:384–422, 1964.

Lombard, E. Le signe de l'elevation de la voix. *Annales des Maladies de L'oreille, du Larynx, du Nez et du Pharynx*, 37(101–119):25, 1911.

Lorenzo-Trueba, J., Barra-Chicote, R., Raitio, T., Obin, N., Alku, P., Yamagishi, J., and Montero, J. M. Towards glottal source controllability in expressive speech synthesis. In *Proceedings of Interspeech*, 2012.

Lu, H., Ling, Z.-H., Lei, M., Wang, C.-C., Zhao, H.-H., Chen, L.-H., Hu, Y., Dai, L.-R., and Wang, R.-H. The USTC system for Blizzard Challenge 2009. In *Proceedings of Blizzard Challenge Workshop*, 2009. URL `http://festvox.org/blizzard`.

Lu, H., Ling, Z.-H., Wei, S., Dai, L.-R., and Wang, R.-H. Automatic error detection for unit selection speech synthesis using log likelihood ratio based SVM classifier. In *Proceedings of Interspeech*, pages 162–165, 2010.

Lu, H., Ling, Z.-H., Dai, L.-R., and Wang, R.-H. Building HMM based unit-selection speech synthesis system using synthetic speech naturalness evaluation score. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5352–5355, 2011.

Lu, H., King, S., and Watts, O. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis. In *Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW8)*, pages 261–265, 2013.

Ma, C., Kamp, Y., and Willems, L. F. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, 12(1):69–81, 1993.

Ma, C., Kamp, Y., and Willems, L. F. A Frobenius norm approach to glottal closure detection from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(2):258–265, 1994.

Maia, R., Toda, T., Zen, H., Nankaku, Y., and Tokuda, K. An excitation model for HMM-based speech synthesis based on residual modeling. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, 2007.

Maia, R., Zen, H., and Gales, M. J. F. Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters. In *Proceedings of the 7th ISCA Speech Synthesis Workshop (SSW7)*, pages 88–93, 2010.

Maia, R., Akamine, M., and Gales, M. J. F. Complex cepstrum as phase information in statistical parametric speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4581–4584, 2012.

Maia, R., Akamine, M., and Gales, M. J. F. Complex cepstrum for statistical parametric speech synthesis. *Speech Communication*, 55(5):606–618, 2013.

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

Malme, C. I. Detectability of small irregularities in a broadband noise spectrum. *Massachutes Institute of Technology, Research Laboratory of Electronics Quarterly Progress Report*, 52:139–141, 1959.

Markel, J. D. and Gray, A. H. *Linear Prediction of Speech*. Springer-Verlag, 2nd edition, 1980.

Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. Voice characteristics conversion for HMM-based speech synthesis system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1611–1614, 1997.

Matausek, M. and Batalov, V. A new approach to the determination of the glottal waveform. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 28(6):616–622, 1980.

Mathieson, L. Normal-disordered continuum. In Kent, R. D. and Ball, M. J., editors, *Voice quality measurement*, pages 3–12. Singular Publishing Group, 2000.

Matsui, K., Pearson, S. D., Hata, K., and Kamai, T. Improving naturalness in text-to-speech synthesis using natural glottal source. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 769–772, Apr. 1991.

Mayo, C., Clark, R. A. J., and King, S. Multidimensional scaling of listener responses to synthetic speech. In *Proceedings of Interspeech*, pages 1725–1728, 2005.

Mayo, C., Clark, R. A. J., and King, S. Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis. *Speech Communication*, 53(3):311–326, 2011.

McCree, A. V. and Barnwell, T. P., III. A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Transactions on Speech and Audio Processing*, 3(4):242–250, 1995.

Meen, D. and Svendsen, T. The NTNU concatenative speech synthesizer. In *Proceedings of Blizzard Challenge Workshop*, 2010. URL `http://festvox.org/blizzard`.

Mehta, D. D., Zañartu, M., Feng, S. W., Cheyne, H. A., and Hillman, R. E. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *IEEE Transactions on Biomedical Engineering*, 59(11):3090–3096, 2012.

Merritt, T., Raitio, T., and King, S. Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis. In *Proceedings of Interspeech*, pages 1509–1503, 2014.

Milenkovic, P. Glottal inverse filtering by joint estimation of an AR system with a linear input model. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 34(1):28–42, 1986.

Miller, R. L. Nature of the vocal cord wave. *Journal of the Acoustical Society of America*, 31(6):667–677, 1959.

Miyanaga, K., Masuko, T., and Kobayashi, T. A style control technique for HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 1437–1439, 2004.

Monsen, R. and Engebretson, A. Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, 62(4):981–993, 1977.

Moore, B. C. J. *An Introduction to the Psychology of Hearing*. Academic Press, 1982.

Moore, B. C. J. Interference effects and phase sensitivity in hearing. *Philosoph-*

ical transactions. Series A, Mathematical, Physical, and Engineering Sciences, 360(1794):833–858, 2002.

Moore, B. C. J. and Glasberg, B. R. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74:750–753, 1983.

Moore, B. C. J. and Glasberg, B. R. Difference limens for phase in normal and hearing-impaired subjects. *Journal of the Acoustical Society of America*, 86:1351–1365, 1989.

Moulines, E. and Di Francesco, R. Detection of the glottal closure by jumps in the statistical properties of the speech signal. *Speech Communication*, 9(5–6):401–418, 1990.

Moulines, E. and Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5–6):453–467, 1990.

Moxley, R. A. On the relationship between speech and writing with implications for behavioral approaches to teaching literacy. *The Analysis of Verbal Behavior*, 8:127–140, 1990.

Murphy, P. J. Perturbation free measurement of the harmonics-to-noise ratio in voice signals using pitch-synchronous harmonic analysis. *Journal of the Acoustical Society of America*, 105:2866–2881, 1999.

Murphy, P. J., McGuigan, K. G., Walsh, M., and Colreavy, M. Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals. *Journal of the Acoustical Society of America*, 123(3):1642–1652, 2008.

Murty, K. and Yegnanarayana, B. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613, 2008.

Muthukumar, P. K., Black, A. W., and Bunnell, H. T. Optimizations and fitting procedures for the Liljencrants-Fant model for statistical parametric speech synthesis. In *Proceedings of Interspeech*, pages 397–401, 2013.

Naylor, P., Kounoudes, A., Gudnason, J., and Brookes, M. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):34–43, 2007.

Ní Chasaide, A., Yanushevskaya, I., and Gobl, C. Voice source dynamics in intonation. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*, pages 1470–1473, 2011.

Nord, L., Ananthapadmanabha, T. V., and Fant, G. Perceptual tests using an interactive source filter model and considerations for synthesis strategies. *Journal of Phonetics*, 14:435–442, 1986.

Norris, J. R. *Markov Chains*. Cambridge University Press, 1998.

Nose, T., Yamagishi, J., Masuko, T., and Kobayashi, T. A style control technique for HMM-based expressive speech synthesis. *IEICE Transactions on Information and Systems*, E90-D(9):1406–1413, 2007.

Odell, J. J. *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, Queens' College, University of Cambridge, Cambridge, UK, 1995.

Ohm, G. S. Ueber die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. *Annalen der Physik und Chemie*, 135(8):513–565, 1843.

Öhman, S. Word and sentence intonation: A quantitative model. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 8(2–3):20–54, 1967.

Okubo, T., Mochizuki, R., and Kobayashi, T. Hybrid voice conversion of unit selection and generation using prosody dependent HMM. *IEICE Transactions on Information and Systems*, E89-D(11):2775–2782, 2006.

Oppenheim, A. and Schafer, R. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, 16(1):221–226, 1968.

O'Shaughnessy, D. *Speech Communications: Human and Machine*. IEEE Press, 2nd edition, 2000.

Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y., and Tokuda, K. Recent development of the HMM-based singing voice synthesis system – Sinsy. In *Proceedings of the 7th ISCA Speech Synthesis Workshop (SSW)*, pages 211–216, 2010.

Oura, K., Zen, H., Nankaku, Y., Lee, A., and Tokuda, K. Tying covariance matrices to reduce the footprint of HMM-based speech synthesis systems. In *Proceedings of Interspeech*, pages 1759–1762, 2009.

Paliwal, K. and Kleijn, W. Quantization of LPC parameters. In Kleijn, W. and Paliwal, K., editors, *Speech Coding and Synthesis*, chapter 12. Elsevier, 1995.

Pantazis, Y., Rosec, O., and Stylianou, Y. On the properties of a time-varying quasi-harmonic model of speech. In *Proceedings of Interspeech*, pages 1044–1047, 2008.

Paris, C. R., Gilson, R. D., Thomas, M. H., and Silver, N. C. Effect of synthetic voice intelligibility on speech comprehension. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(2):335–340, 1995.

Patterson, R. D. A pulse ribbon model of monaural phase perception. *Journal of the Acoustical Society of America*, 82(5):1560–1586, 1987.

Peterson, G. E. and Barney, H. L. Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, 24:175–184, 1952.

Pfitzinger, H. R. Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. In Hoffmann, R. and Mixdorff, H., editors, *Proceedings of the 3rd International Conference on Speech Prosody*, pages 6–9. TUDpress, Dresden, 2006.

Pickett, J. M. Effects of vocal force on the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, 28(5):902–905, 1956.

Pickett, J. M. *The Acoustics of Speech Communication. Fundamentals, Speech Perception Theory, and Technology*. Allyn and Bacon, 1999.

Pinto, N., Childers, D., and Lalwani, A. Formant speech synthesis: Improving production quality. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 37(12):1870–1887, 1989.

Plomp, R. and Mimpen, A. Improving the reliability of testing the speech reception threshold of sentence. *Audiology*, 18(1):43–52, 1979.

Plomp, R. and Steeneken, H. J. M. Effect of phase on the timbre of complex tones. *Journal of the Acoustical Society of America*, 46(2B):409–421, 1969.

Plumpe, M., Acero, A., Hon, H.-W., and Huang, X.-D. HMM-based smoothing for concatenative speech synthesis. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 2751–2754, 1998.

Pobloth, H. and Kleijn, W. B. On phase perception in speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 29–32, 1999.

Pollet, V. and Breen, A. Synthesis by generation and concatenation of multiform segments. In *Proceedings of Interspeech*, pages 1825–1828, 2008.

Pols, L. C. W. Assessment of text-to-speech synthesis systems. In Fourcin, A. J., Harland, G., Barry, W., and Hazan, V., editors, *Speech Input and Output Assessment. Multilingual Methods and Standards*, chapter III, pages 53–81, 251–266. Ellis Horwood Ltd, Chichester, UK, 1989.

Poritz, A. Linear predictive hidden Markov models and the speech signal. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 7, pages 1291–1294, 1982.

Pulakka, H. Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography. Master's thesis, Department of Signal Processing and Acoustics, Helsinki University of Technology, Finland, 2005.

Qian, Y., Yan, Z.-J., Wu, Y.-J., Soong, F. K., Zhang, G., and Wang, L. An HMM trajectory tiling (HTT) approach to high quality TTS – Microsoft entry to Blizzard Challenge 2010. In *Proceedings of Blizzard Challenge Workshop*, 2010. URL `http://festvox.org/blizzard`.

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

Rabiner, L. and Juang, B.-H. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

Rabiner, L. and Schafer, R. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NY, 1978.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. HMM-based Finnish text-to-speech system utilizing glottal inverse filtering. In *Proceedings of Interspeech*, pages 1881–1884, 2008.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. Comparison of formant enhancement methods for HMM-based speech synthesis. In *Proceedings of the 7th ISCA Speech Synthesis Workshop (SSW7)*, pages 334–339, 2010.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4564–4567, 2011a.

Raitio, T., Suni, A., Vainio, M., and Alku, P. Analysis of HMM-based Lombard speech synthesis. In *Proceedings of Interspeech*, pages 2781–2784, 2011b.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1): 153–165, 2011c.

Raitio, T., Takanen, M., Santala, O., Suni, A., Vainio, M., and Alku, P. On measuring the intelligibility of synthetic speech in noise – Do we need a realistic noise environment? In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4025–4028, 2012a.

Raitio, T., Kane, J., Drugman, T., and Gobl, C. HMM-based synthesis of creaky voice. In *Proceedings of Interspeech*, pages 2316–2320, 2013a.

Raitio, T., Suni, A., Vainio, M., and Alku, P. Comparing glottal-flow-excited statistical parametric speech synthesis methods. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7830–7834, 2013b.

Raitio, T., Lu, H., Kane, J., Suni, A., Vainio, M., King, S., and Alku, P. Voice source modelling using deep neural networks for statistical parametric speech synthesis. In *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, 2014a.

Raitio, T., Suni, A., Juvela, L., Vainio, M., and Alku, P. Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort. In *Proceedings of Interspeech*, pages 1969–1973, 2014b.

Raitio, T., Suni, A., Vainio, M., and Alku, P. Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise. *Computer Speech and Language*, 28(2):648–664, 2014c.

Raitio, T. Hidden Markov model based Finnish text-to-speech system utilizing glottal inverse filtering. Master's thesis, Department of Signal Processing and Acoustics, Helsinki University of Technology, Finland, 2008.

Raitio, T., Suni, A., Vainio, M., and Alku, P. Wideband parametric speech synthesis using warped linear prediction. In *Proceedings of Interspeech*, 2012b.

Raitio, T., Suni, A., Pohjalainen, J., Airaksinen, M., Vainio, M., and Alku, P. Analysis and synthesis of shouted speech. In *Proceedings of Interspeech*, pages 1544–1548, 2013c.

Rajkumar, R., White, M., Speer, S. R., and Ito, K. Evaluating prosody in synthetic speech with online (eye-tracking) and offline (rating) methods. In *Proceedings of the 7th ISCA Speech Synthesis Workshop (SSW7)*, pages 276–281, 2010.

Rao, K. and Yegnanarayana, B. Prosody modification using instants of significant excitation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):972–980, 2006.

Rao, K., Prasanna, S., and Yegnanarayana, B. Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. *IEEE Signal Processing Letters*, 14(10):762–765, 2007.

Riegelsberger, E. and Krishnamurthy, A. Glottal source estimation: Methods of applying the LF-model to inverse filtering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 542–545, 1993.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 749–752, 2001.

Rix, A. W., Hollier, M. P., Beerends, J. G., and Hekstra, A. P. PESQ – The new ITU standard for end-to-end speech quality assessment. In *Proceedings of the 109th Convention of the Audio Engineering Society*, 2000.

Roberts, G. O. and Smith, A. F. M. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49:207–216, 1994.

Roebel, A. and Rodet, X. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *Proceedings of Digital Audio Effects (DAFx)*, volume 8, pages 30–35, 2005.

Roebel, A., Villavicencio, F., and Rodet, X. On cepstral and all-pole based spectral envelope modeling with unknown model order. *Pattern Recognition Letters*, 28(11):1343–1350, 2007.

Rosenberg, A. Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 49(2B):583–590, 1971.

Rosenberg, A. E., Schafer, R. W., and Rabiner, L. R. Effects of smoothing and quantizing the parameters of formant-coded voiced speech. *Journal of the Acoustical Society of America*, 50(6B):1532–1538, 1971.

Rossing, T. D., Moore, F. R., and Wheeler, P. A. *The Science of Sound*. Addison-Wesley, San Francisco, CA, USA, 3rd edition, 2002.

Rostolland, D. Acoustic features of shouted voice. *Acustica*, 50(2):118–125, 1982a.

Rostolland, D. Phonetic structure of shouted voice. *Acustica*, 51(2):80–89, 1982b.

Rostolland, D. Intelligibility of shouted voice. *Acustica*, 57(3):103–121, 1985.

Rothenberg, M. Glottal noise during speech. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology (KTH), Stockholm*, 15(2–3):1–10, 1974.

Rothenberg, M. Acoustic interaction between the glottal source and the vocal tract. In Stevens, K. and Hirano, M., editors, *Vocal fold physiology*, pages 305–323. University of Tokyo Press, 1981.

Rothenberg, M., Carlson, R., Granström, B., and Lindqvist-Gauffin, J. A three parameter model for the glottal source. In Fant, G., editor, *Speech Communication*, volume 2, pages 235–243. Almqvist and Wiksell, Stockholm, 1975.

Rouibia, S. and Rosec, O. Unit selection for speech synthesis based on a new acoustic target cost. In *Proceedings of Interspeech*, pages 2565–2568, 2005.

Russell, M. and Moore, R. Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 10, pages 5–8, 1985.

Saino, K., Zen, H., Nankaku, Y., Lee, A., and Tokuda, K. HMM-based singing voice synthesis system. In *Proceedings of Interspeech*, pages 2274–2277, 2006.

Saon, G., Soltau, H., Nahamoo, D., and Picheny, M. Speaker adaptation of neural network acoustic models using i-vectors. In *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 55–59, 2013.

Sapienza, C., Stathopoulos, E., and Dromey, C. Approximations of open quotient and speed quotient from glottal airflow and EGG waveforms: Effects of measurement criteria and sound pressure level. *Journal of Voice*, 12(1): 31–43, 1998.

Saratxaga, I., Erro, D., Hernáez, I., Sainz, I., and Navas, E. Use of harmonic phase information for polarity detection in speech signals. In *Proceedings of Interspeech*, pages 1075–1078, 2009.

Scherer, K. R. Vocal communication of emotion: A review of research paradigms.

Speech Communication, 40(1–2):227–256, 2003.

Schnell, K. Estimation of glottal closure instances from speech signals by weighted nonlinear prediction. In Chetouani, M., Hussain, A., Gas, B., Milgram, M., and Zarader, J.-L., editors, *Advances in Nonlinear Speech Processing*, volume 4885 of *Lecture Notes in Computer Science*, pages 221–229. Springer Berlin Heidelberg, 2007.

Schroeder, M. R. and Strube, H. W. Flat-spectrum speech. *Journal of the Acoustical Society of America*, 79(5):1580–1583, 1986.

Schroeder, M. R. New results concerning monaural phase sensitivity. *Journal of the Acoustical Society of America*, 31(11):1579–1579, 1959.

Segi, H., Takagi, T., and Ito, T. A concatenative speech synthesis method using context dependent phoneme sequences with variable length as search units. In *Proceedings of the 5th ISCA Workshop on Speech Synthesis (SSW5)*, pages 115–120, 2004.

Shannon, M., Zen, H., and Byrne, W. Autoregressive models for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(3):587–597, 2013.

Shichiri, K., Sawabe, A., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. Eigenvoices for HMM-based speech synthesis. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 1269–1272, 2002.

Shinoda, K. and Watanabe, T. MDL-based context-dependent subword modeling for speech recognition. *Journal of the Acoustical Society of Japan (E)*, 21(2):79–86, 2000.

Silén, H., Helander, E., Nurminen, J., and Gabbouj, M. Ways to implement global variance in statistical speech synthesis. In *Proceedings of Interspeech*, 2012.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. ToBI: A standard for labeling English prosody. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 867–870, 1992.

Simple4All. Simple4All project website, 2014. URL `http://simple4all.org`.

Sluijter, A. M. C., van Heuven, V. J., and Pacilly, J. J. A. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1):503–513, 1997.

Smith, A. F. M. and Roberts, G. O. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 55:3–23, 1993.

Smits, R. and Yegnanarayana, B. Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing*, 3(5):325–333, 1995.

Sonntag, G. P. and Portele, T. PURR – A method for prosody evaluation and investigation. *Computer Speech and Language*, 12(4):437–451, 1998.

Soong, F. K. and Juang, B.-H. Line spectrum pair (LSP) and speech data compression. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 9, pages 37–40, 1984.

Sorin, A., Shechtman, S., and Pollet, V. Uniform speech parameterization for multi-form segment synthesis. In *Proceedings of Interspeech*, pages 337–340, 2011.

Sorin, A., Shechtman, S., and Pollet, V. Refined inter-segment joining in multi-form speech synthesis. In *Proceedings of Interspeech*, pages 790–794, 2014.

Stan, A. and Giurgiu, M. A superpositional model applied to F0 parameterization using DCT for text-to-speech synthesis. In *Proceedings of the 6th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6, 2011.

Steeneken, H. J. M. and Houtgast, T. A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America*, 67 (1):318–326, 1980.

Stevens, S. S., Volkmann, J., and Newman, E. B. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

Stewart, J. Q. An electrical analogue of the vocal organs. *Nature*, 110:311–312, 2012.

Story, B. H. An overview of the physiology, physics and modeling of the sound source for vowels. *Acoustical Science and Technology*, 23(4):195–206, 2002.

Story, B. H., Titze, I. R., and Hoffman, E. A. Vocal tract shapes and area functions from magnetic resonance imaging (MRI). *Journal of the Acoustical Society of America*, 98(5):2930–2930, 1995.

Strik, H. Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *Journal of the Acoustical Society of America*, 103(5):2659–2669, 1998.

Strik, H. and Boves, L. On the relation between voice source parameters and prosodic features in connected speech. *Speech Communication*, 11(2–3):167–174, 1992.

Strik, H., Cranen, B., and Boves, L. Fitting a LF-model to inverse filter signals. In *Proceedings of Eurospeech*, pages 103–106, 1993.

Strube, H. Determination of the instant of glottal closure from the speech wave. *Journal of the Acoustical Society of America*, 56(5):1625–1629, 1974.

Strube, H. W. Linear prediction on a warped frequency scale. *Journal of the Acoustical Society of America*, 68(4):1071–1076, 1980.

Sturmel, N., d'Alessandro, C., and Doval, B. A comparative evaluation of the zeros of Z transform representation for voice source estimation. In *Proceedings of Interspeech*, pages 558–561, 2007.

Sturmel, N., d'Alessandro, C., and Rigaud, F. Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4517–4520, 2009.

Stylianou, Y. Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 377–380, 1999.

Stylianou, Y. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1): 21–29, 2001.

Stylianou, Y., Cappe, O., and Moulines, E. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*,

6(2):131–142, 1998.

Sulter, A. and Wit, H. Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age. *Journal of the Acoustical Society of America*, 100(5):3360–3373, 1996.

Summers, W. V., Pisoni, D., Bernacki, R., Pedlow, R., and Stokes, M. Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, 84(3):917–928, 1988.

Sundberg, J., Titze, I., and Scherer, R. Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source. *Journal of Voice*, 7:15–29, 1993.

Sundberg, J., Andersson, M., and Hultqvist, C. Effects of subglottal pressure on professional baritone singers' voice sources. *Journal of the Acoustical Society of America*, 105(2):1965–1971, 1999a.

Sundberg, J., Cleveland, T., Stone, R., and Iwarsson, J. Voice source characteristics in six premier country singers. *Journal of Voice*, 13(1):168–183, 1999b.

Sung, J., Hong, D., Oh, K., and Kim, N. Excitation modeling based on waveform interpolation for HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 813–816, 2010.

Suni, A., Raitio, T., Vainio, M., and Alku, P. The GlottHMM speech synthesis entry for Blizzard Challenge 2010. In *Proceedings of Blizzard Challenge Workshop*, 2010. URL `http://festvox.org/blizzard`.

Suni, A., Raitio, T., Vainio, M., and Alku, P. The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation. In *Proceedings of Blizzard Challenge Workshop*, 2011. URL `http://festvox.org/blizzard`.

Suni, A., Raitio, T., Vainio, M., and Alku, P. The GlottHMM entry for Blizzard Challenge 2012 – Hybrid approach. In *Proceedings of Blizzard Challenge Workshop*, 2012. URL `http://festvox.org/blizzard`.

Suni, A., Aalto, D., Raitio, T., Alku, P., and Vainio, M. Wavelets for intonation modeling in HMM speech synthesis. In *Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW8)*, pages 285–290, 2013.

Suni, A., Raitio, T., Gowda, D., Karhila, R., Gibson, M., and Watts, O. The Simple4All entry to the Blizzard Challenge 2014. In *Proceedings of Blizzard Challenge Workshop*, 2014. URL `http://festvox.org/blizzard`.

Swietojanski, P. and Renals, S. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, 2014.

't Hart, J. Discriminability of the size of pitch movements in speech. *IPO Annual Progress Report*, 9:56–63, 1974.

Taal, C. H., Hendriks, R. C., Heusdens, R., Jensen, J., and Kjems, U. An evaluation of objective quality measures for speech intelligibility prediction. In *Proceedings of Interspeech*, pages 1947–1950, 2009.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transactions on Information and Systems*, E88-D(11):2484–2491, 2005.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. A style adaptation technique for speech synthesis using HSMM and suprasegmental features. *IEICE Transactions on Information and Systems*, E89-D(3): 1092–1099, 2006.

Takamichi, S., Toda, T., Neubig, G., Sakti, S., and Nakamura, S. A postfilter to modify the modulation spectrum in HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 290–294, 2014.

Talkin, D. Voicing epoch determination with dynamic programming. *Journal of the Acoustical Society of America*, 85(S1):S149–S149, 1989.

Talkin, D. A robust algorithm for pitch tracking (RAPT). In Kleijn, W. B. and Paliwal, K. K., editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier, 1995.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 805–808, 2001.

Taylor, P. Unifying unit selection and hidden Markov model speech synthesis. In *Proceedings of Interspeech*, pages 1758–1761, 2006.

Taylor, P. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

Teager, H. M. and Teager, S. M. Evidence for nonlinear sound production mechanisms in the vocal tract. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 241–261. Kluwer Academic Publishers, 1990.

Teutenberg, J., Watson, C., and Riddle, P. Modelling and synthesising F0 contours with the discrete cosine transform. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3973–3976, 2008.

Thomas, M., Gudnason, J., and Naylor, P. Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):82–91, 2012.

Thomas, M. R. P., Gudnason, J., and Naylor, P. A. Data-driven voice source waveform modelling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3965–3968, 2009.

Tierney, L. Markov chains for exploring posterior distributions, with discussion. *The Annals of Statistics*, 22:1701–1762, 1994.

Timcke, R., von Leden, H., and Moore, P. Laryngeal vibrations: Measurements of the glottic wave. *Archives of Otolaryngology*, 68:1–19, 1958.

Tiomkin, S., Malah, D., Shechtman, S., and Kons, Z. A hybrid text-to-speech system that combines concatenative and statistical synthesis units. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1278–1288, 2011.

Titze, I. and Sundberg, J. Vocal intensity in speakers and singers. *Journal of the Acoustical Society of America*, 91(5):2936–2946, 1992.

Titze, I. R. *Principles of Voice Production*. Prentice Hall, 1994.

Toda, T. and Tokuda, K. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, E90-D(5):816–824, 2007.

Toda, T., Black, A. W., and Tokuda, K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.

Tokuda, K., Kobayashi, T., and Imai, S. Speech parameter generation from HMM using dynamic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 660–663, 1995a.

Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., and Imai, S. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In *Proceedings of Eurospeech*, volume 1, pages 757–760, 1995b.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 229–232, 1999.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1315–1318, 2000.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems*, E85-D(3):455–464, 2002a.

Tokuda, K., Zen, H., and Black, A. W. An HMM-based speech synthesis system applied to English. In *Proceedings of the 2002 IEEE Workshop on Speech Synthesis*, pages 227–230, 2002b.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K. Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013.

Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. Mel-generalized cepstral analysis – A unified approach to speech spectral estimation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 1043–1046, 1994.

Tooher, M. and McKenna, J. G. Variation of glottal LF parameters across F0, vowels, and phonetic environment. In *Voice Quality: Functions, Analysis and Synthesis, ISCA Tutorial and Research Workshop (VOQUAL '03)*, pages 41–46, 2003.

Traunmüller, H. and Eriksson, A. Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107(6):3438–3451, 2000.

Tuan, V. N. and d'Alessandro, C. Robust glottal closure detection using the wavelet transform. In *Proceedings of Eurospeech*, pages 2805–2808, 1999.

Tüske, Z., Golik, P., Schlüter, R., and Ney, H. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In *Proceedings of Inter-*

speech, pages 890–894, 2014.

Umeda, N., Matsui, E., Suzuki, T., and Omura, H. Synthesis of fairy tales using an analog vocal tract. *Proceedings of the 6th International Congress on Acoustics*, pages B159–162, 1968.

Vainio, M., Järvikivi, J., Werner, S., Volk, N., and Välikangas, J. Effect of prosodic naturalness on segmental acceptability in synthetic speech. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, pages 143–146, 2002.

Vainio, M., Suni, A., Järveläinen, H., Järvikivi, J., and Mattila, V.-V. Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish. *Journal of the Acoustical Society of America*, 118(3):1742–1750, 2005.

Vainio, M., Airas, M., Järvikivi, J., and Alku, P. Laryngeal voice quality in the expression of focus. In *Proceedings of Interspeech*, pages 921–924, 2010.

Vainio, M., Suni, A., Raitio, T., Nurminen, J., Järvikivi, J., and Alku, P. New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis. In *Proceedings of Interspeech*, pages 1703–1706, 2009.

Valentini-Botinhao, C., Yamagishi, J., and King, S. Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5112–5115, 2011a.

Valentini-Botinhao, C., Yamagishi, J., and King, S. Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise? In *Proceedings of Interspeech*, pages 1837–1840, 2011b.

van den Berg, J., Zantema, J. T., and Doornenbal, P. On the air resistance and the Bernoulli effect of the human larynx. *Journal of the Acoustical Society of America*, 29(5):626–631, 1957.

Veeneman, D. and BeMent, S. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 33:369–377, 1985.

Veldhuis, R. and Klabbers, E. On the computation of the Kullback–Leibler measure for spectral distances. *IEEE Transactions on Speech and Audio Processing*, 11(1):100–103, 2003.

Vilkman, E., Lauri, E.-R., Alku, P., Sala, E., and Sihvo, M. Loading changes in time based parameters of glottal flow waveforms in different ergonomic conditions. *Folia Phoniatrica et Logopaedica*, 49:247–263, 1997.

Vincent, D., Rosec, O., and Chonavel, T. A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 525–528, 2007.

Vishnubhotla, S., Fernandez, R., and Ramabhadran, B. An autoencoder neural-network based low-dimensionality approach to excitation modeling for HMM-based text-to-speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4614–4617, 2010.

Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2): 260–269, 1967.

Voiers, W. D. Diagnostic evaluation of speech intelligibility. In Hawley, M. E., editor, *Speech Intelligibility and Speaker Recognition*, pages 374–387. Dowden, Hutchinson and Ross, Stroudsburg, PA, 1977.

von Helmholtz, H. L. F. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Vieweg und Sohn, Braunschweig, 1863.

von Kempelen, W. R. *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. J. B. Degen, Wien, 1791.

Walker, J. and Murphy, P. Advanced methods for glottal wave extraction. In Faundez-Zanuy, M. et al., editors, *Nonlinear Analyses and Algorithms for Speech Processing*, pages 139–149. Springer Berlin/Heidelberg, 2005.

Watts, O., Stan, A., Mamiya, Y., Suni, A., Burgos, J. M., and Montero, J. M. The Simple4All entry to the Blizzard Challenge 2013. In *Proceedings of Blizzard Challenge Workshop*, 2013. URL `http://festvox.org/blizzard`.

Watts, O. *Unsupervised Learning for Text-to-Speech Synthesis*. PhD thesis, University of Edinburgh, 2012.

Wong, D., Markel, J., and Gray Jr., A. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Audio, Speech, and Language Processing*, 27(4):350–355, 1979.

Wouters, J. and Macon, M. Unit fusion for concatenative speech synthesis. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 302–305, 2000.

Wu, Y.-J. and Wang, R.-H. Minimum generation error training for HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 89–92, 2006.

Yamagishi, J. *Average-Voice-Based Speech Synthesis*. PhD thesis, Tokyo Institute of Technology, Japan, 2006.

Yamagishi, J. and King, S. Simple methods for improving speaker-similarity of HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4610–4613, 2010.

Yamagishi, J. and Kobayashi, T. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Transactions on Information and Systems*, E90-D(2):533–543, 2007.

Yamagishi, J., Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. A training method of average voice model for HMM-based speech synthesis. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E86-A(8):1956–1963, 2003.

Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, E88-D(3):503–09, 2005.

Yamagishi, J., Nose, T., Zen, H., Toda, T., and Tokuda, K. Performance evaluation of the speaker-independent HMM-based speech synthesis system "HTS 2007" for the Blizzard Challenge 2007. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3957–3960, 2008.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):66–83, 2009a.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., and Renals, S. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17 (6):1208–1230, 2009b.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Guan, Y., Hu, R., Oura, K., Wu, Y.-J., Tokuda, K., Karhila, R., and Kurimo, M. Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):984–1004, 2010.

Yamagishi, J., Veaux, C., King, S., and Renals, S. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, 33(1):1–5, 2012.

Yang, J.-H., Zhao, Z.-W., Jiang, Y., Hu, G.-P., and Wu, X.-R. Multi-tier non-uniform unit selection for corpus-based speech synthesis. In *Proceedings of Blizzard Challenge Workshop*, 2006. URL `http://festvox.org/blizzard`.

Yanushevskaya, I., Gobl, C., Kane, J., and Ní Chasaide, A. An exploration of voice source correlates of focus. In *Proceedings of Interspeech*, pages 462–465, 2010.

Yao, K., Yu, D., Seide, F., Su, H., Deng, L., and Gong, Y. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 366–369, 2012.

Yegnanarayana, B. and Gangashetty, S. Epoch-based analysis of speech signals. *Sadhana*, 36(5):651–697, 2011.

Yegnanarayana, B. and Veldhuis, N. Extraction of vocal-tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing*, 6:313–327, 1998.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. Speaker interpolation in HMM-based speech synthesis system. In *Proceedings of Eurospeech*, pages 2523–2526, 1997.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. Duration modeling for HMM-based speech synthesis. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 29–32, 1998.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of Eurospeech*, pages 2374–2350, 1999.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. Mixed excitation for HMM-based speech synthesis. In *Proceedings of Eurospeech*, pages 2259–2262, 2001.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.-Y., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. The Hidden Markov Model Toolkit (HTK) Version 3.4. Software, 2006. URL `http://htk.eng.cam.ac.uk`.

Young, S. J., Odell, J. J., and Woodland, P. C. Tree-based state tying for high

accuracy acoustic modelling. In *Proceedings of the Workshop on Human Language Technology*, pages 307–312, 1994.

Yu, D., Yao, K., Su, H., Li, G., and Seide, F. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7893–7897, 2013a.

Yu, J., Zhang, M., Tao, J., and Wang, X. A novel HMM-based TTS system using both continuous HMMs and discrete HMMs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 709–712, 2007.

Yu, K. and Young, S. Continuous F0 modeling for HMM based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1071–1079, 2011.

Yu, Y., Zhu, F., Li, X., Liu, Y., Zou, J., Yang, Y., Yang, G., Fan, Z., and Wu, X. Overview of SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013. In *Proceedings of Blizzard Challenge Workshop*, 2013b. URL `http://festvox.org/blizzard`.

Zen, H. and Toda, T. An overview of nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Proceedings of Blizzard Challenge Workshop*, 2005. URL `http://festvox.org/blizzard`.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, pages 294–299, 2007.

Zen, H., Senior, A., and Schuster, M. Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7962–7966, 2013.

Zen, H. and Senior, A. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3872–3876, 2014.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. Hidden semi-Markov model based speech synthesis. In *Proceedings of Interspeech*, pages 1393–1396, 2004.

Zen, H., Tokuda, K., and Black, A. W. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

Zhang, Z., Xian, X., Luo, L., and Wu, X. PKU Mandarin speech synthesis system for Blizzard 2009. In *Proceedings of Blizzard Challenge Workshop*, 2009. URL `http://festvox.org/blizzard`.

Zwicker, E. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, 33(2):248–248, 1961.

Errata

Publication IV

The sentence in Section 2.2. “However, with a large number of concatenation points, the full Viterbi search is not computationally feasible.” is not correct, and later a Viterbi search was implemented.

Speech is the most natural way through which humans communicate, and today speech synthesis is utilised in various applications. However, the performance of modern speech synthesisers falls far short from the abilities of human speakers—synthesising intelligible and natural sounding speech with desired contextual and speaker characteristics and appropriate speaking style is extremely difficult. This thesis aims to improve both the naturalness and expressivity of speech synthesis by proposing new methods for voice source modelling in statistical parametric speech synthesis. With accurate estimation and appropriate modelling of the voice source signal, which is known to be the origin for several essential acoustic cues in spoken communication, various expressive voices are created with high degree of naturalness and intelligibility. Evaluations in various listening contexts show that speech created with the proposed methods is assessed to be more suitable than that generated with current techniques, thus providing potentially large benefits in many speech synthesis applications.

Aalto-DD 40/2015

ISBN 978-952-60-6136-8 (printed)
ISBN 978-952-60-6137-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

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