Helsinki University of Technology
Department of Signal Processing and Acoustics
Espoo 2008 Report 3

METHODS AND STUDIES OF LARYNGEAL VOICE QUALITY ANALYSIS IN SPEECH PRODUCTION

Matti Airas

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission for public examination and debate in Auditorium S1, Faculty of Electronics, Communications and Automation, Helsinki University of Technology, Espoo, Finland, on the 23rd of May 2008, at 12 o'clock noon.

Helsinki University of Technology

Faculty of Electronics, Communications and Automation

Department of Signal Processing and Acoustics

Teknillinen korkeakoulu

Elektroniikan, tietoliikenteen ja automaation tiedekunta

Signaalinkäsittelyn ja akustiikan laitos

Helsinki University of Technology
Department of Signal Processing and Acoustics
P.O. Box 3000
FI-02015 TKK
Tel. +358 9 4511
Fax +358 9 460224
E-mail Lea.Soderman@tkk.fi

ISBN 978-951-22-9385-8
ISSN 1797-4267

Espoo, Finland 2008

ABSTRACT OF DOCTORAL DISSERTATION	HELSINKI UNIVERSITY OF TECHNOLOGY P. O. BOX 1000, FI-02015 TKK http://www.tkk.fi
Author Matti Airas	
Name of the dissertation Methods and Studies of Laryngeal Voice Quality Analysis in Speech Production	
Manuscript submitted Feb. 4, 2008	
Date of the defence May 23, 2008	
□ Monograph	☒ Article dissertation (summary + original articles)

Faculty	Faculty of Electronics, Communications and Automation
Department	Department of Signal Processing and Acoustics
Field of research	Speech Processing
Opponent(s)	Prof. Federico Avanzini
Supervisor	Prof. Paavo Alku
Instructor	

Abstract

Voice quality, defined by John Laver as the characteristic auditory colouring of a speaker's voice, is a significant feature of speech, and it is used to signal various properties such as emotions, intentions, and mood of the speaker. While voice quality measurement techniques and algorithms have been developed, much work is needed to obtain a comprehensive view of the function and analysis of human voice in the production of different voice qualities. Two major research questions are presented in this thesis: First, how can the most important laryngeal voice quality features be analyzed, and second, how do the voice quality features affect different facets of vocal expression? To answer these questions, five separate studies of the analysis methodology and two studies regarding the voice quality behaviour were published. The methodology articles describe a voice source analysis software package; a comparison of multiple voice source parameters in breathy, normal, and pressed phonation; a method for evaluating inverse filtering algorithms; comparison of two inverse filtering algorithms; and a method for analyzing intensity regulation of speech. One analysis article studies changes in the laryngeal voice quality when different emotions are expressed in speech and another voice quality changes in expression of prominence in continuous speech. The methodology studies resulted in new tools, methods, and guidelines for voice source analysis, while the analysis studies provide information on how voice quality is used in expressive speech.

Keywords speech processing, voice quality, glottal inverse filtering, vocal expression of emotions	
ISBN (printed) 978-951-22-9385-8	ISSN (printed) 1797-4267
ISBN (pdf) 978-951-22-9386-5	ISSN (pdf) 1797-4275
Language English	Number of pages 176
Publisher Helsinki University of Technology, Department of Signal Processing and Acoustics	
Print distribution Report 3 / TKK, Department of Signal Processing and Acoustics	
☒ The dissertation can be read at http://lib.tkk.fi/Diss/	

VÄITÖSKIRJAN TIIVISTELMÄ

TEKNILLINEN KORKEAKOULU
PL 1000, 02015 TKK
http://www.tkk.fi

Tekijä Matti Airas

Väitöskirjan nimi
Kurkunpäässä tuotetun puheäänen laadun analyysimenetelmiä ja tutkimuksia

Käsikirjoituksen päivämäärä 4.2.2008

Väitöstilaisuuden ajankohta 23.5.2008

□ Monografia ☒ Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)

Tiedekunta	Elektroniikan, tietoliikenteen ja automaation tiedekunta
Laitos	Signaalinkäsittelyn ja akustiikan laitos
Tutkimusala	Puheenkäsittely
Vastaväittäjä(t)	Prof. Federico Avanzini
Työn valvoja	Prof. Paavo Alku
Työn ohjaaja	

Tiivistelmä

Puheäänen laatu eli John Laverin määritelmän mukainen puhujan äänen luonteenomainen sävy on puheen merkittävä piirre. Sitä käytetään viestittämään puhujan eri tiloja kuten emootioita, aikeita ja mielialoja. Vaikka monia puheäänen laadun mittausmenetelmiä ja -algoritmeja on kehitetty, kattavan kuvan hankkiminen ihmisen puheentuoton ja äänenlaadun toiminnasta vaatii vielä paljon työtä. Tässä väitöskirjassa esitetään puheenäänen laadun tutkimuksesta kaksi keskeistä kysymystä: ensinnäkin, kuinka tärkeimpiä kurkunpäässä tuotettuja puheäänen laadun piirteitä voidaan analysoida, ja toiseksi, kuinka puheääen laadun piirteet vaikuttavat puheilmaisun eri aspekteihin? Jotta näihin kysymyksiin voitiin vastata, tässä väitöskirjatyössä tehtiin viisi julkaisua puheäänen laadun analyysimenetelmistä ja kaksi puheäänen laadun käyttäytymisestä. Menetelmäartikkelit kuvasivat kehitettyä äänilähdeanalyysiohjelmistoa, usean äänilähdeparametrin vertailua vuotoisassa, normaalissa ja puristeisessa äännössä, käänteissuodatusalgoritmien arviointimenetelmää, kahden eri käänteissuodatusmenetelmän vertailua sekä puheen intensiteettisäätelyn analyysimenetelmää. Toinen puheäänenlaadun muutoksia käsitelleistä artikkeleista kuvasi puheäänen laadunvaihteluita eri emootioita ilmaistaessa, kun taas toisessa tutkittiin puheäänen laadun muutoksia ilmaistaessa painoa jatkuvassa puheessa. Menetelmätutkimusten tuloksena syntyi uusia ohjelmistoja, analyysimenetelmiä ja ohjenuoria äänilähteen analyysiin, kun taas äänenlaadun muutoksia käsitelleet julkaisut tuottivat tietoa äänilähteen ja puheäänen laadun käytöstä ekspressiivisen puheen aikana.

Asiasanat digitaalinen signaalinkäsittely, puheenkäsittely, puheen äänenlaatu, puheen emootioilmaisu

ISBN (painettu)	978-951-22-9385-8	ISSN (painettu)	1797-4267
ISBN (pdf)	978-951-22-9386-5	ISSN (pdf)	1797-4275
Kieli	Englanti	Sivumäärä	176

Julkaisija Teknillinen korkeakoulu, Signaalinkäsittelyn ja akustiikan laitos

Painetun väitöskirjan jakelu Raportti 3 / TKK:n Signaalinkäsittelyn ja akustiikan laitos

☒ Luettavissa verkossa osoitteessa http://lib.tkk.fi/Diss/

To the people within my heart

Preface

As an adolescent I aspired to lasting fame, I craved factual certainty, and I thirsted for a meaningful vision of human life – so I became a scientist. This is like becoming an archbishop so you can meet girls.
– Matt Cartmill

This thesis work was conceived in late 2002, when I was (inefficiently) finishing my Master's Thesis. I was having a fit of my Peter Pan syndrome, and did not know what to do when I grow up. Those were the circumstances when I began working with Prof. Alku. Several years have passed, and I am now finishing my Doctoral Thesis. I am still suffering of the Peter Pan syndrome, but at least I now have an idea what I will be doing.

This thesis would not have been possible without funding by Academy of Finland project numbers 200859 and 111848 (Research Programme on Neuroscience, NEURO) and a graduate school position in Graduate School of Language Technology in Finland. Additional support has been provided by *Kaupallisten ja teknillisten tieteiden tukisäätiö KAUTE* and *Emil Aaltosen Säätiö*. All this financial support is highly acknowledged.

For finishing this thesis, I am indebted to Paavo Alku. Not only did he hire me in the first place, but he believed in me even in those times when I myself did not. He has been able to coach me through the ordeals of my doctoral research by thoughtful application of both the carrot and the stick, and is, in my opinion, as good an academic supervisor as one could hope for. Actually, never mind that—he has also provided me with friendship, which I value even more.

I have had an opportunity to work with several extraordinary personalities at the Laboratory of Acoustics. Tom Bäckström, Hannu Pulakka, Eva Björkner, and Laura Lehto have provided me both countless hours of professional inspiration and unprofessional entertainment. In addition to them, I would like to thank my other co-authors Brad Story, Johan Sundberg, and Martti Vainio for their contribution in realizing this thesis. Furthermore, the pre-examiners Peter Murphy and Erkki Vilkman provided invaluable commentary to the manuscript, while Luis R. de Jussilainen Costa patiently corrected my English speling. Lea Söderman has helped with practical matters throughout the years I have been working at the laboratory. She has a huge heart, and has been a great friend off-duty as well. Moreover, my pets would not have survived my trips without her!

The brilliant people in the increasingly Kafkaesque setting of the Finnish university environment make for an interesting place to work in. I am a bit hesitant to publicly admit how many hours I have wasted with my office-mate and friend Jyri Pakarinen discussing the multiverse, the essence of the scientific method, and similar issues (not that many, actually), and matters of the more immature kind (a lot, to say the least). The same applies to Heikku, whose pure and innocent mind we have been able to taint to a confusing degree. I could carry on, but suffice it to say: thank you, Carlo (In Memoriam 1980–2008), Cumi, Hanna, Henkka, Jussi, Laura Enflo, Mara, Matti, Miikka, Pete, Riitta, Seppo, Toni, Tuomas, Unski, Ville, and all the others I forgot to mention!

Last, I would like to express my warmest gratitude to my parents Kaija and Kalervo, sisters Laura and Katri, and their children Emma, Sarah, Vili, Eemeli, and Kalle. They make up a great family, which I feel warm and proud to belong to. Thank you.

Matti Airas
Otaniemi, April 30th, 2008

Contents

List of Publications

I Matti Airas and Paavo Alku, “Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient.” *Phonetica*, 63(1), pp. 26–46, March 2006.

II Matti Airas, Paavo Alku, and Martti Vainio, “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 in Biomedical Applications (MAVEBA 2007)*, pp. 135–138, Florence, Italy, December 13–15, 2007.

III Matti Airas, “TKK Aparat: An Environment for Voice Inverse Filtering and Parameterization.” *Logopedics Phoniatrics Vocology*, 33(1), pp. 49–64, 2008.

IV Matti Airas and Paavo Alku, “Comparison of Multiple Voice Source Parameters in Different Phonation Types.” In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, pp. 1410–1413, Antwerpen, Belgium, August 27–31, 2007.

V Paavo Alku, Brad Story, and Matti Airas, “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(2), pp. 102–113, 2006.

VI Laura Lehto, Matti Airas, Eva Björkner, Johan Sundberg, and Paavo Alku, “Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types.”, *Journal of Voice*, 21(2), pp. 138–150, March 2007.

VII Paavo Alku, Matti Airas, Eva Björkner, and Johan Sundberg, “An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity.” *Journal of the Acoustical Society of America*, 120(2), pp. 1052–1062, August 2006.

Author's contribution

Publication I: "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient"

The acquisition of data, including the selection of emotions, elicitation method, and recitation material, as well as the actual recording process was performed by the co-author of the publication and other members of the joint project. The present author segmented and analyzed the data, performed the statistical analyses, and wrote the manuscript. Paavo Alku provided comments to the manuscript.

Publication II: "Laryngeal voice quality changes in expression of prominence in continuous speech"

The experimental design was done in collaboration with the publication co-authors. The present author participated in recording the data, performed the analysis, and wrote the bulk of the manuscript. The co-authors contributed to the introduction and to the conclusions.

Publication III: "TKK Aparat: An Environment for Voice Inverse Filtering and Parameterization"

The vast majority of the code was written by the present author. Tom Bäckström and Hannu Pulakka contributed code to some sections. All the experiments and analyses were design and performed by the present author. The author also prepared the manuscript, to which Paavo Alku contributed some comments.

Publication IV: "Comparison of Multiple Voice Source Parameters in Different Phonation Types"

The present author designed the experimental setup, participated in the recording of the data, analyzed the recordings, and wrote the first manuscript version. The co-author of the publication provided comments to the experimental setup and to the manuscript.

Publication V: "Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production"

The author of the thesis presented an earlier version of the manuscript at a conference and contributed comments on the manuscript.

Publication VI: "Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types"

The present author participated in the experimental design, was one of the participants in the inverse filtering phase, and performed the statistical analyses of the study. The description of the statistical analyses and the results section were also written by the present author.

Publication VII: "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity"

The present author participated in the experimental design and interpretation of the results of the study, as well as performed the statistical analyses and created all of the figures in the manuscript.

List of Abbreviations and Symbols

A_{ac}	Glottal flow peak-to-peak amplitude
ANOVA	Analysis of variance
AQ	Amplitude quotient
ARMA	Autoregressive moving average
CCD	Charge-coupled device
CMOS	Complementary metal oxide semiconductor
ClQ	Closing quotient
d_{min}	Minimum value of the glottal flow derivative
DAP	Discrete all-pole modelling (El-Jaroudi and Makhoul, 1991)
$\Delta\mathrm{H}_{12}$	Harmonic level difference (also: H1-H2)
DIF	Direct inverse filtering
EGG	Electroglottography
f_0	Fundamental frequency
FE	Finite element
FEM	Finite element method
HRF	Harmonic richness factor
IAIF	Iterative adaptive inverse filtering
LF	Liljencrants-Fant (Fant et al., 1985)
LP	Linear prediction
MRI	Magnetic resonance imaging
NAQ	Normalized amplitude quotient
OQ	Open quotient
PGG	Photoglottography
PSP	Parabolic spectrum parameter (Alku and Vilkman, 1996a)

SPL	Sound pressure level
SQ	Speed quotient
T	Length of the glottal flow cycle
T_{c}	Length of the glottal flow closing phase
T_{o}	Length of the glottal flow open phase
T_{op}	Length of the glottal flow opening phase
VKG	Videokymography

List of Figures

1 Introduction

“It’s not what you say, but how you say it.” The age-old adage about the form vs. function dichotomy condenses the essence of this thesis. A complex and rich language is one of the distinct features of the human race. Human beings, as profoundly social animals, are also, in addition to the explicit, verbal channel of vocal communication, exceptionally talented in picking up the tone of speech and evaluating the speaker’s message, intentions, mood, and even health status from it. This tone of speech, called the *voice quality*, constitutes much of the second, implicit channel of vocal communication. Laver (1980) defines it as “the characteristic auditory colouring of a speaker’s voice”. Although the definition uses perceptual aspects of voice, voice quality is firmly rooted in the physiological process of voice production, which also is the viewpoint of this thesis.

In traditional speech sciences, the existence and significance of the voice quality in vocal communication is acknowledged, but often ignored. In disciplines such as phonetics, research of vocal expression of emotions, logopedics, and vocology, measures only indirectly related to the voice quality are customarily utilized. Measures such as the fundamental frequency, intensity, and spectral tilt reflect more profound phenomena in the voice production system, which, due to methodological and conceptual deficiencies, have traditionally gained insufficient attention. Voice quality changes, directly related to the actual physiological source of voiced speech, should be able to describe and explain the secondary parameter variations at a more basic level. Therefore, direct analysis of the voice source should give a better insight on the factors actually causing the observed voice quality changes.

Within the last two or three decades, the methodology and actual study of the voice quality and the voice production have gained increasing attention. Measurement techniques and algorithms have been developed, and considerable knowledge of the details of the voice apparatus has been acquired. Still, much work needs to be done to obtain a comprehensive view of the function and analysis of human voice in production of different voice qualities.

Most of the studies utilize a voice source separation method called inverse filtering (Miller, 1959; Fant, 1960). All proposed inverse filtering methods make assumptions regarding the separability of the source and the filter, even though those assumptions are known to be invalid in real life. Due to this discrepancy and other possible imperfections caused by the inverse filtering methods, evaluation of the validity of the acquired estimates of the airflow between the vocal folds (the glottal volume velocity waveform, or glottal flow) should be essential in the field. However, since direct measurements of the glottal flow are not practically possible, direct evaluation of the inverse filtering methods is not possible. Hence, most studies have relied on simple qualitative inspection of the acquired waveform shapes or the spectrum as the verification method. Another issue related to inverse filtering methodology is that few comparative studies of different inverse filtering methods have been performed. Therefore, little data exists on the discrepancies of the glottal flow estimates acquired using different algorithms, and the effect of the inverse filtering method on the acquired flow waveforms cannot be easily estimated.

Quantitative analysis of the estimated glottal flow waveforms requires parameterization of the results. While numerous parameters have been suggested, little is known about their relative differences, their performance in measuring different features of the voice source, or how the parameter values can be compared to each other.

In contrast to the relative abundance of the different published voice source analysis and parameterization methods, surprisingly few software packages for those purposes are publicly available. The lack of existing research and development platforms has created an unnecessary barrier of entry to the field, with every research group having to implement their own algorithms at a great time expense and a risk of implementation errors.

Due to practical reasons, comparatively few studies of voice source on continuous speech have been performed. Instead, stationary vowels or very short, fixed utterances are commonly used. As marked differences can be found between stationary vowels and continuous speech, future research needs to be able to address the problems associated with the analysis of continuous speech and to ascertain whether knowledge acquired using stationary phonation is applicable to natural speech.

Studies utilizing voice source analysis techniques have already been performed in various fields such as occupational voice (Vilkman et al., 1999; Lehto, 2007), classical singing (Björkner, 2006), pathological voice evaluation (Murphy, 2006), and perception of emotions (Gobl and Ní Chasaide, 2003). However, improved knowledge on the voice source functioning and its relationship to the voice quality could be useful in, for example, reliable recognition of vocal emotions. Emotion recognition and detection techniques could be used, for example, in improving human-computer interaction, automated telephone services, and automatic labelling of recordings. Furthermore, the contemporary high-quality methods for speech synthesis mostly allow for limited voice quality changes with reduced naturalness. The expressivity of speech synthesis techniques could be dramatically improved if voice quality adjustments could be performed. This would be a valuable feature in areas such as human-computer interaction both in traditional and mobile computing environments, as well as in the gaming and film industries.

1.1 Scope of this thesis

Two research questions were presented in this study:

- How can the most important laryngeal voice quality features be analyzed?
- How do the voice quality features affect different facets of vocal expression?

The first research question concerns the voice quality analysis methods. Some practical limitations were set for the methodology from the beginning. First, the analysis methods should be non-invasive and sufficiently discrete in nature, so that the actual data acquisition would not overtly affect the tasks at hand. For example, the equipment and the environment required for the use of high-speed imaging of the vocal folds would hardly allow for realistic expression of emotions, with the exception of fear and anxiety! In practice, these limitations dictated the use of microphone recordings and voice inverse filtering throughout the study. Yet, the validity and the reliability of the inverse filtering methods, the data parameterization techniques, as well as the actual working environment were discussed to shed light on the question. The goal was to establish a suitable methodology for voice quality studies.

The second research question concerns the actual gathering of knowledge regarding the voice quality features. As a complete examination of all different voice quality dimensions would have been far beyond the scope of a single doctoral thesis, the present studies were limited to the breathy–pressed (or lax–tense) axis of laryngeal voice quality. To answer the second research question, the relationship of different aspects of vocal expression such as vocal expression of emotions and prominence to the pressedness were examined. Hence, the goal was to obtain data on the voice source behaviour in voice quality changes in continuous speech.

1.2 Organization of the thesis

The doctoral thesis consists of this summary and seven publications. Five of the articles were published in international peer-reviewed journals and two in reviewed conference proceedings. The Articles III–VII discuss the voice source analysis and parameterization methods, while the Articles I and II address the actual voice quality differences in different contexts.

The summary provides a broader overview and background information on the topics and concepts of the thesis at a more detailed level than in the articles. In section 2, the human voice production mechanism is described. Section 3 discusses different methods of artificially modelling the voice production. Section 4 gives a treatment on different voice analysis techniques, while section 5 summarizes different voice source parameterization methods. The concepts of voice quality and vocal expression of emotion are treated in section 6. Section 7 summarizes the articles and section 8 concludes the thesis.

2 Voice production system

2.1 Subglottal structures

The process of producing voiced sounds starts from the airflow caused by the pressure exerted on the lungs. Lungs are a pair of organs located in the thorax. Their primary function is respiration by absorption of oxygen from air and excretion of carbon dioxide from the bloodstream. However, due to that function, they also act as powerful bellows, facilitating airstream through the respiratory airways (Rossing, 1990). The total volume of the lungs is about 6–7 litres, of which two litres is residual volume, which can never be depleted unless the lungs collapse (Titze, 1994). Thus, the vital capacity of lungs is 4–5 litres. In low levels of physical activity, only 10–15% of the vital capacity is inhaled and exhaled.

In the expiratory process, the elastic recoil of the lungs and ribs is sufficient to initiate exhaling (Titze, 1994). In the first phase of expiration, the diaphragm is used to regulate the elastic recoil to control the lung pressure. In the latter expiratory phases of active breathing, exhalation is further continued by a pressure exerted on the lungs by the internal intercostal and the abdominal muscles (West, 2000). The muscles induce an overpressure, which in average phonation is 0.5–1 kPa (Hirano, 1981). The pressure differential between the lungs and the ambient air causes an airflow through the respiratory airways. This airflow has been measured to be about 0.07–0.20 l/s in sustained phonation, with great individual variation (Hirano, 1981).

The air from the lungs erupts through the respiratory airways, the lower part of which consist of the bronchi and the trachea. Unlike the upper part, consisting of the larynx and the vocal tract discussed below, the trachea is a fairly rigid, 10–12 cm long tubular structure, about 25 mm in diameter (Stevens, 2000).

2.2 Larynx

The larynx, illustrated in the cross-section of the vocal organs in Fig. 2.1, is situated in the neck, where it is surrounded by a large number of blood vessels, nerves, glands, and other supply lines of the human body (Titze, 1994). These impose severe space constraints on the larynx, which affect the larynx's operation. Furthermore, the larynx is used in several auxiliary functions as well. For example, in swallowing the larynx is moved upwards so that the airway can be sealed. In yawning the pharynx expands and the larynx is depressed to widen the airway. The range of vertical movement in these operations can be several centimeters. The larynx can also move forward when greater air flow is needed or when a lump of food is swallowed. The need for mobility of the laryngeal structure precludes any rigid bony attachments to the skeleton, and so most of the laryngeal framework is in the form of cartilages.

The different cartilages of the larynx are shown in Fig. 2.2. The *thyroid cartilage* shields the inner structures and forms a rigid structure in front of them (Titze, 1994). It comprises of two plates that are joined at the midline at an angle of about 90° to 120°. The angle is usually smaller in adult males than in women and children, resulting in a prominence known as the Adam's apple. At the posterior borders of the thyroid cartilage, four projections arise. The two upward projections, *superior cornu*, connect the thyroid cartilage via a ligament to the hyoid bone. The downward projections, *inferior cornu*, join with the cricoid cartilage. The *cricoid cartilage* lies directly below the thyroid cartilage, forming a solid ring completely surrounding the laryngeal airway. It may be thought of as the most superior tracheal ring, but is different in shape and construction. The two *arytenoid cartilages* are situated on top of the posterior portion

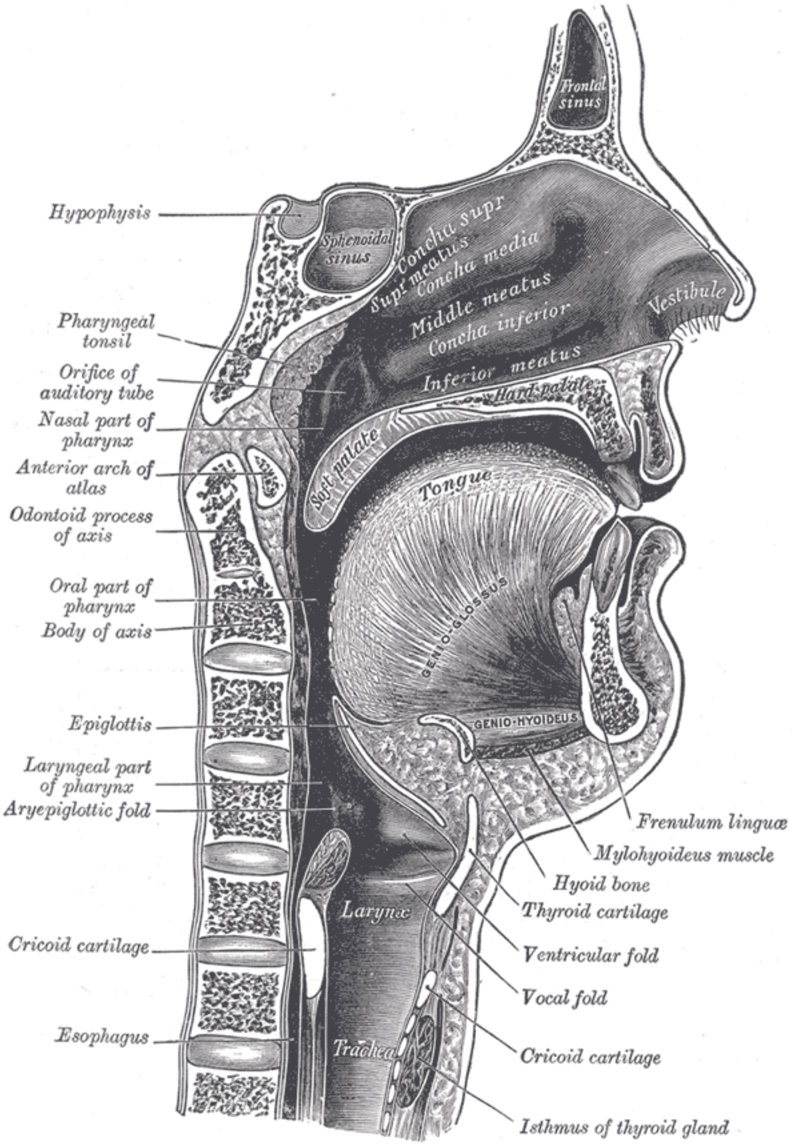


Figure 2.1: A sagittal cross-section of the respiratory airways, illustrating the key elements of the voice production system (Gray and Lewis, 1918).

of the cricoid cartilage. The vocal ligament attaches to the anterior projection at the base of the arytenoid cartilage. The arytenoid cartilages are readily movable due to the highly flexible cricoarytenoid joint. The arytenoid cartilages can move not only in the medial-lateral direction, but they also can rock in the anterior-posterior direction. The *epiglottis* is a lid-like cartilage at the top of the larynx, which folds over the entryway to the larynx when tight closure of the airway is needed.

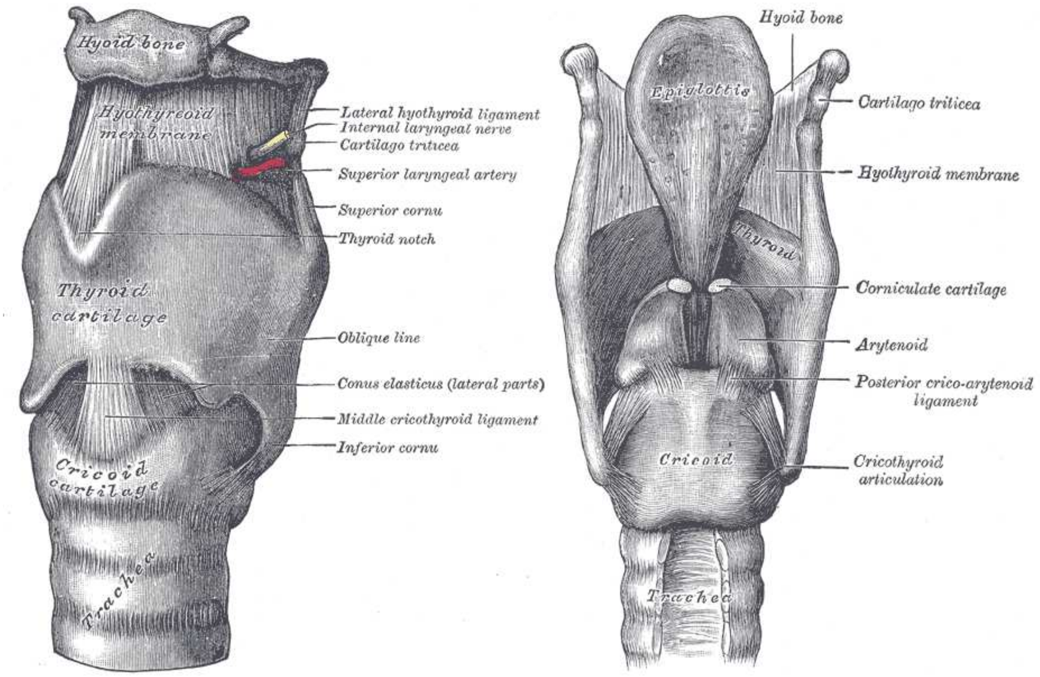


Figure 2.2: Antero-lateral and posterior views of the ligaments and cartilages of the larynx (Gray and Lewis, 1918).

The hyoid bone is the only bone related to the larynx (Titze, 1994). While not part of the larynx proper, it connects to the thyroid cartilage through the thyroid membrane and the superior cornu. Many muscles are also anchored to this bone.

The muscles of the larynx may be divided into two groups, the intrinsics and the extrinsics (Titze, 1994). The intrinsic muscles interconnect the different cartilages of the larynx, while the extrinsic muscles connect the laryngeal structures to its surroundings, such as the sternum or the hyoid bone. The two *thyroarytenoid muscles*, shown in Fig. 2.3, are connected to the thyroid and arytenoid cartilages. They make up the bulk of the vocal folds. When contracted, they pull the arytenoid cartilages forward, thereby shortening, thickening and stiffening the vocal folds. The two *cricothyroid muscles* connect the cricoid and thyroid cartilages. They elevate the cricoid arch and depress the thyroid lamina, thereby shortening the cricothyroid space and lengthening the vocal folds. Thus, they are the primary pitch-control muscles. The *lateral and posterior cricoarytenoid* and *interarytenoid muscles* connect the arytenoids to the cricoid and to each other, thus facilitating the movements of the arytenoids. The different extrinsic laryngeal muscles control the vertical movements of the larynx and the hyoid bone.

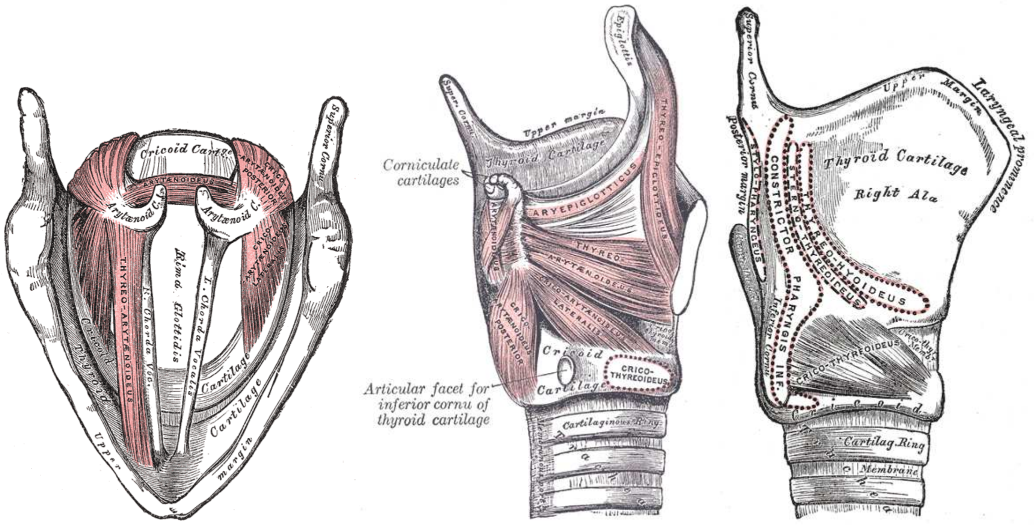


Figure 2.3: Top and two side views of larynx muscles and their attachments to the cartilages (Gray and Lewis, 1918).

The vocal folds are located at the narrowest portion of the airway in the larynx. They are a valve-like structure composed of rather thick and flexible mucosa and muscle tissue. In breathing, they are used in controlling the airflow and sealing off the trachea from the pharynx when needed. The length of the vocal fold is about 1.6 cm in males and 1 cm in females, and their mass is approximately 1 gram (Hirano et al., 1981). The aperture between the vocal folds is called the *glottis*. In males, the average glottal peak cross-sectional areas of 126 mm^2 during inspiration and 27 mm^2 during phonation have been measured (Brancatisano et al., 1983; Hertegård and Gauffin, 1995). A schematic of a coronal section of a vocal fold is shown in Fig. 2.4. The outermost layer is a thin skin made up of layered and scale-like epithelium (Hirano, 1981). It is 0.05–0.10 mm thick, and acts as a capsule whose purpose is to maintain the shape of the vocal fold. The lamina propria resides between the epithelium and the thyroarytenoid muscle. It is a layered system of non-muscular tissues that can be divided into three layers: superficial, intermediate, and deep. The superficial layer consists of elastic protein fibers surrounded by interstitial fluids. It can be likened to a mass of soft gelatin. The intermediate layer also consists primarily of elastic fibers, although in this layer they are mainly longitudinally aligned. Furthermore, mixed with the elastic fibers there are also some inextensible collagen fibers. The deep layer consists of longitudinally aligned collagenous, inextensible fibers. The superficial layer is approximately 0.5 mm thick in the middle of the vocal fold, while the intermediate and deep layers together are about 1–2 mm. The thyroarytenoid muscle forms the major portion of the vocal fold, being approximately 7–8 mm thick (Titze, 1994). The structure of the vocal fold, however, is not uniform along its length. The intermediate layer thickens and the superficial layer becomes thinner towards the ends of the vocal folds, and there also exist masses of elastic fibers—an extension of the intermediate layer—at both ends of the vocal folds, cushioning the membranes' attachments to the surrounding cartilages (Hirano, 1981).

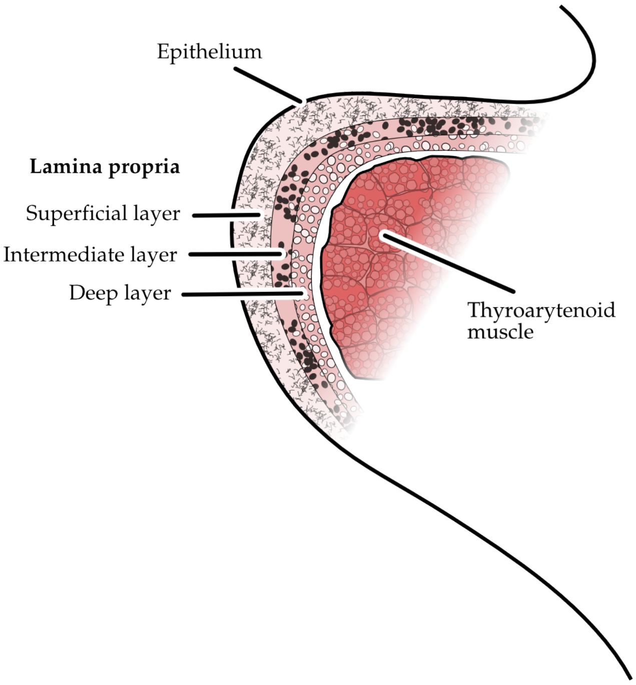


Figure 2.4: Schematic drawing of a coronal section through the right vocal fold. (After Titze 1994.)

2.3 Vocal tract

The respiratory airway above the larynx is called the vocal tract. The vocal tract consists of the pharynx, mouth, and the nasal cavity (see Fig. 2.1). The shape of the vocal tract can be altered by moving the tongue vertically and longitudinally. The tongue placement affects the relative diameters of different parts of the vocal tract, thus affecting the vocal tract resonances, i.e., the formants. In non-nasalized phonation, the soft palate seals the nasal cavity from the pharynx, so that the airflow only occurs through the pharyngeal cavity and the mouth. Furthermore, the length of the vocal tract may be adjusted by moving the larynx vertically or by widening or constricting the mouth opening.

The average vocal tract length of young adult males has been found to be 15.5 cm and of females 13.9 cm (Fitch and Giedd, 1999). The vocal tract volume has been reported to vary, depending on the individual, vowel, and voice quality, between 19 to 80 cm^3 (Stevens, 2000; Story et al., 2001).

2.4 Principles of vocal fold oscillation

According to the traditional myoelastic-aerodynamic theory of vocal fold oscillation, vocal folds vibrate due to the Bernoulli forces caused by the sufficiently high-velocity airstream through the glottis pulling the vocal folds together (Titze, 1994). After the closure of the glottis, the elastic properties of the tissue together with the increased subglottal pressure force the separation of the vocal folds. The movement outwards then continues, until the elastic forces of the tissue limit and reverse it.

Unfortunately, the myoelastic-aerodynamic theory is unable to explain self-sustained oscillation. The Bernoulli forces are equal in strength in both the inward and outward motion of the vocal folds, and thus no net energy would be accrued to sustain the vibration (Titze, 1994). The separatory forces exerted by the pressure difference over the glottis, for one, require relatively complete closure of the glottis, and would not explain how the vibration can occur even without it.

The airstream through the glottis does not vanish in thin air; instead, it travels through the vocal tract. The moving column of air in the vocal tract also has a certain mass, and hence, inertia. The inertia of the air in the vocal tract provides positive feedback to the vocal fold vibration: when the folds are opening, the air column is accelerated by the airflow below, causing a positive pressure at the vocal folds, which further pushes them apart. Respectively, when the folds are closing, the inertia of the air column causes a negative pressure at the glottis, further pulling the vocal folds together (Titze, 1994). Since the vocal tract coupling is related to the cyclic nature of the vocal fold oscillation, it provides a positive feedback and introduces an asymmetry between the vocal fold motion and the driving forces, thus establishing sufficient conditions for self-sustained oscillation of the vocal folds.

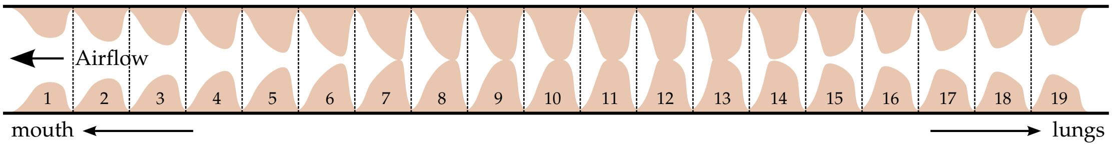


Figure 2.5: Schematic time-series of the vocal fold vibration. Frames 1–6 illustrate the closing phase. The glottis is closed during frames 7–13, during which the contact area moves upwards, until the folds separate again in frames 14–19. Throughout the cycle, the lower portion of the vocal folds (on the right-hand side in the illustration) leads the movement.

Observations of vocal fold oscillations have shown that the vocal fold movement is not uniform in the medial plane of the vocal folds (Titze, 1994; Timcke et al., 1958; van den Berg et al., 1957). Instead, it has been noticed that the vocal fold moves with a wavelike motion—the upper portion of the vocal fold lags the lower portion, as illustrated in figure 2.5. This nonuniform movement appears to exhibit important properties which, in addition to the vocal tract coupling, are essential in sustaining the oscillation. Due to the vocal fold shape during the cycle, a pressure difference exists between the inward and outward movement, and as in the case of vocal tract coupling, the pressure difference is aligned to amplify the oscillations.

3 Modelling of voice production

Due to ethical and methodological problems, validation of the voice production phenomena *in vivo* may be impractical or even impossible. Therefore, modelling of voice production has been widely used in studies of the voice source. Physical modelling forms either a descriptive model such as a mass-spring model, or a model that is physically as precise as possible, like the finite element method (FEM). In addition, it is possible to use acoustical modelling, which focuses on the properties of the major acoustic signals instead of the physiological details of the voice production system. Acoustic modelling methods utilize, in one form or another, the source filter theory, in which the glottal airflow is represented using some approximate mathematical formula, and the vocal tract is regarded as a filter according to Fant's source-filter model (Fant, 1960).

3.1 Physical modelling of voice source

3.1.1 Mass-spring models

Mass-spring models are a popular approach for simulating vocal fold behavior. Mass-spring models are constructed by lumping the vocal fold mass into few discrete mass elements, which are connected to each other and to a rigid boundary with springs and damping elements. Figure 3.1 shows three different mass-spring models of varying complexity.

In a one-mass model of the vocal folds, the vocal folds are considered a simple second-order system consisting of a mass, a spring, and a damper (Flanagan and Landgraf,

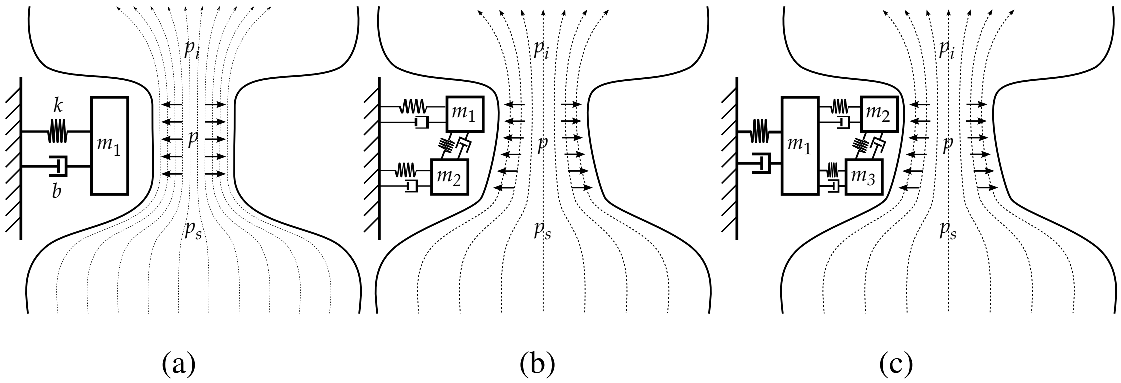


Figure 3.1: The one-, two-, and three-mass models of the vocal folds.

1968). The mass represents the total mass of the vocal folds. The spring constant represents the tissue elasticity, while the damping accounts for the energy losses in the vibration caused by tissue viscosity. The spring constant can be varied according to vocal fold tension and the damping can be determined experimentally.

The one-mass model of vocal fold vibration is able to represent only the lateral displacement of the vocal folds. As the mucosal wave cannot be modelled, a one-mass model is capable of self-sustaining oscillation only in the presence of a vocal tract and the inertance of the air column within it (Titze, 1994). Thus, one-mass models exaggerate the effect of coupling of vocal tract and the vocal folds and are generally not sufficient for research purposes.

To alleviate the observed shortcomings of the one-mass model, two-mass models of the vocal folds were introduced in a widely cited paper by Ishizaka and Flanagan (1972). An example of such a model is illustrated in Fig. 3.1 (b). The two-mass model is able to simulate the phase difference between the upper and lower parts of the vocal folds, described in Sec. 2.4. The two degrees of freedom allow the mucosal wave to be represented in addition to the overall tissue displacement. Thus, a two-mass model facilitates oscillation even without the presence of a coupled vocal tract inertance (Story, 2002).

A limitation of the two-mass models is that their discretization of tissue does not capture the layered structure of the vocal folds (Story, 2002). Although the lower mass in the model presented by Ishizaka and Flanagan (1972) is made thicker and heavier in an effort to simulate the effects of the body layer, the arrangement does not allow for a coupled oscillation of both the cover and body layers. In essence, the two-mass model is a “cover” model rather than a “cover-body” model. Furthermore, in the human vocal folds, stiffness is controlled by the contraction of the thyroarytenoid muscles. In the two-mass model of Ishizaka and Flanagan (1972), there exists no direct physiological correlation between the spring stiffnesses and the effects of muscle contractions (Story and Titze, 1995). These limitations can be overcome by adding another mass to the model, creating a three-mass model (Story and Titze, 1995), shown in Fig. 3.1. The added mass element acts as a body mass, and is positioned laterally to the the two cover masses. The cover mass connections represent the stiffness of the cover tissue as well as the effective coupling stiffness between the body and cover. The body mass is connected to a rigid boundary, and this connection represents the effective stiffness of the body, which depends on the level of contraction of the muscle tissue. This model thus allows for physiologically realistic control parameters characterizing the cover and body tissue features. For example, the contraction of the thyroarytenoid muscle increases the stiffness of the body but may not necessarily stiffen the cover.

Perpetual increase of model complexity may lead to overparameterization of the problem. Even a two-mass model requires setting of as many as 19 parameters to account for phenomena such as nonlinear elastic forces and collisions of the vocal folds, resulting in high computational loads and problems in tuning the parameters. Therefore, Avanzini et al. (2001) and Avanzini (2008) have proposed a modified one-mass model, which retains the simplicity of the traditional one-mass models while still allowing for mimicking of the features of the more complex two-mass models. While their model incorporates a time delay parameter which is not directly physiologically motivated, they gain a high degree of control and low computational requirements in the model.

3.1.2 Finite element method (FEM) models

The lumped-mass models of the vocal folds described in the previous section are useful in describing the main features of the vocal fold vibration. They are, however, insufficient to accurately simulate the tissue dynamics. Their spatial resolution is insufficient to reflect the scale of vocal fold physiology, and the parameters have few direct implications for vocal-fold tissue physiology (Gunter, 2003). FEM is a technique used for solving partial differential equations approximately. In FEM, the modelled entity is tesselated into a triangular mesh, and then the partial differential equations are discretized, describing the properties of the entity to these triangles. The method can be used both for two- and three-dimensional modelling.

Finite element models were first applied to modelling of vocal fold vibration by Alipour and Titze (1985). They discretized the vocal folds into areas representing the body, the cover, and the ligament. To optimize the speed of computation, a 2D/3D hybrid model was used. A coronal section of their vocal fold model is shown in Fig. 3.2

Three-dimensional FE models of the vocal folds, which have a close resemblance to the actual vocal fold physiology, have also been developed. For example, Gunter (2003) developed a vocal fold model with a high spatial resolution for the estimation of effects of pathological and surgical alterations of the vocal folds.

Another approach to FE modelling is to model the fluid flow within the vocal tract. For example, Hannukainen et al. (2007) modelled the spectral properties of a stationary vocal tract using magnetic resonance imaging (MRI). Their analysis resulted in

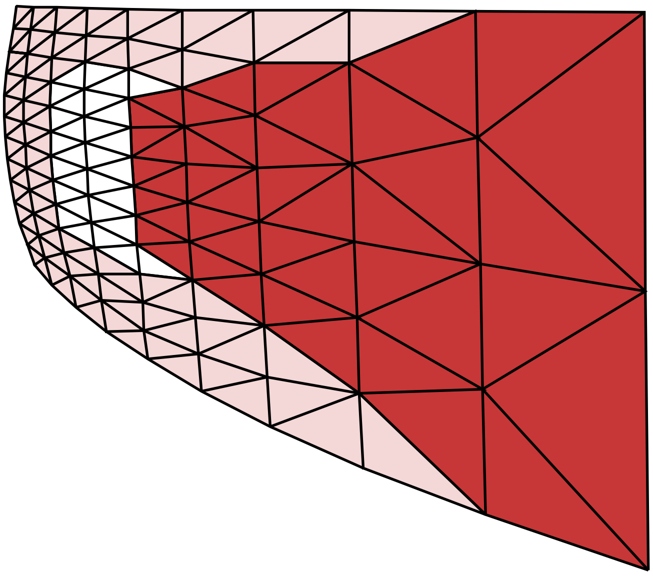


Figure 3.2: A coronal slice of the vocal fold mesh used in the FEM models of Alipour and Titze (1985) and Alipour (2000). The model differentiates between the body, ligament, and cover layers of the vocal folds.

precise vocal tract filters, which then may be applied to source-filter models of speech production, described in the next section. Dedouch et al. (2002) have modelled the production of several Czech vowels using FE models of vocal tract shapes acquired using MRI analysis and a mass-spring model of vocal folds. The formant frequencies computed using the model were in a good agreement with the data on the formant frequencies published in literature.

A combined viscoelastic-acoustic FE model of the voice production system would be a logical continuation of the separate FE models of the vocal folds and the vocal tract. However, the computational complexity issues in combined models have so far limited such implementations.

3.2 Source-filter model

In addition to physical modelling described earlier, another approach is to estimate the acoustic waveform directly, without paying attention to the physiological details of the voice production. In this approach, human voice production is assumed to conform to the source-filter model, in which voiced speech is modelled by three separate processes considered to be linear: the glottal excitation (source), vocal tract filtering, and the lip-radiation effect (Fant, 1960). A block diagram of the source-filter model is shown in Figure 3.3. In the source-filter model of voice production, the voice source acts as as source signal, which is then filtered by the formant filter of the vocal tract as well as the lip radiation filter. This may be represented as an equivalent acoustic circuit, in which the vocal tract in non-nasalized vowels constitutes a distributed series inductance and parallel capacitance per unit length of the vocal tract (Fant, 1973). Since the Laplace transform of such a system only has poles in the negative half-plane, its stability is guaranteed and it is also inversible. Furthermore, since the blocks of the source-filter model are considered linear and to have no interaction, they can be modelled and manipulated with very modest computational requirements. The actual physiological voice production mechanism does not fully conform to Fant's acoustic theory of speech production, as acoustic interaction between the voice source and the vocal tract has been long known to exist. In practice, however, the source-filter model still has been found to be applicable to a multitude of problems.

One of the most common methods of voice source modelling is to construct a synthetic mathematical approximation of the glottal airflow and use the acquired waveform as the source. The method allows for easy modification of the source signal, within the limits of the mathematical model. Numerous source models have been proposed. For example, Rosenberg (1971) presented several alternative simulated glottal flow pulses, most of which were composed of sinusoidal sections. Fant et al. (1985) refined their

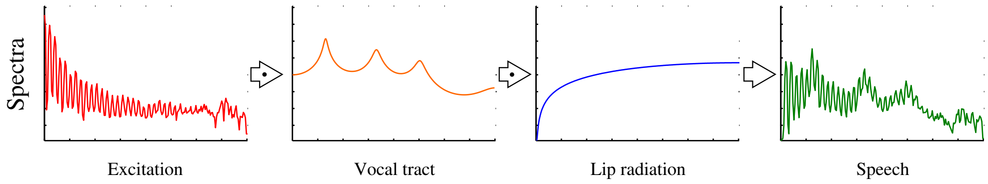
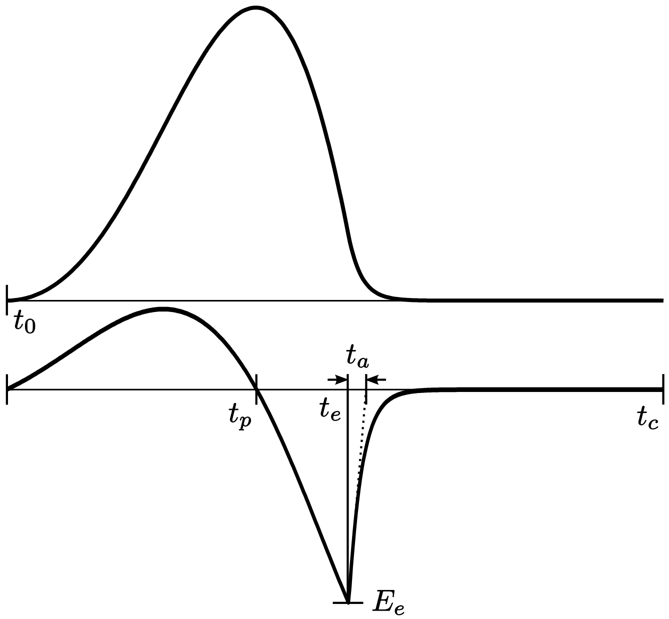


Figure 3.3: A block diagram of the source-filter model of speech production. According to the source-filter theorem, speech production can be split into several independent and therefore linearly separable processes.

previous glottal flow models and proposed the Liljencrants-Fant (LF) model, which consists of exponential and exponential-sinusoidal sections. The waveforms and definition of the LF-model are shown in Fig. 3.4. The flow pulse is defined by five different parameters. Furthermore, Fant (1995) defines five more parameters which are used interchangeably with the original ones. The LF-model is by far the most commonly used synthetic model of the glottal flow, even though the calculation of the waveform is somewhat complicated when compared to most other models. More recently, Veldhuis (1998) has proposed the R++ model, which is a computationally efficient glottal flow model equivalent to the LF-model. Furthermore, Doval et al. (2006) have presented a unified glottal flow model, which is claimed to be spectrally equivalent, and therefore exchangable, with any of the aforementioned models.

In synthesis of vowels using source-filter modelling, the vocal tract filter represents the vowel formant structure. The actual filter parameters may be acquired from analysis of natural speech or synthesized using prior knowledge of the formant locations and bandwidths. The lip-radiation effect corresponds to changing the volume velocity waveform at the lips to a free-field speech pressure signal at a certain distance from the speaker. It can be modelled by a simple differentiation filter. Furthermore, instead of using synthetic glottal flow pulses, glottal flow waveforms acquired using inverse filtering, described in section 4.3, may be used in the synthesis of vowels (Alku et al.,



$$E(t) = E_0 e^{\alpha t} \sin w_g t \qquad (t < t_e)$$

$$E(t) = \frac{-Ee}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] \qquad (t_e < t < t_c)$$

Figure 3.4: The LF-model of glottal flow. The upper graph shows the glottal flow, while the lower is the differentiated glottal flow, acquired by the shown equations. The parameters α, ω_g, and ε are set using boundary conditions described in the original paper (Fant et al., 1985).

1999). This semi-synthetic speech generation method allows for highly natural vowel sounds at the expense of lost flexibility in modifying the pulse shape.

Due to its versatility, the source-filter model of speech production has been applied to a multitude of speech processing tasks, including speech analysis, synthesis, and coding. In speech analysis, important techniques such as inverse filtering, discussed in section 4.3, are based on Fant's model of speech production. In speech synthesis, separated speech production models are applied when versatility and easy modification of the voice quality features of speech are required. In speech coding, methods such as the linear predictive coding conform to the source-filter model, even though there are differences between the two. Specifically, linear predictive coding lumps the spectral structure of the source, filter, and the lip radiation effect together, leaving just a train of impulses as the excitation signal.

4 Voice production analysis techniques

The concealed location of the vocal folds poses difficulties for the analysis of their function. Due to their placement, direct measurement of their vibration is intrusive and difficult. Various methods for their observation have been developed. The observation techniques can generally be divided to three categories. First, imaging techniques concentrate on visual analysis of larynx by observing the vocal folds using a mirror, optic fiber, or a photodiode. These techniques return a visual image of the glottis, a part thereof, or a dynamic brightness variation curve. Second, electrical and electromagnetical glottography methods extract specific features of the vocal fold vibration related to the changing electrical properties of the tissue. Third, acoustical methods estimate the airflow through the glottis by using techniques called inverse filtering, or variations thereof.

4.1 Imaging techniques

Laryngoscopy is a method for visually observing the vocal folds, either using a solid endoscope, or a flexible fiberscope. The fiberscope, illustrated on the left hand side in Fig. 4.1, is inserted through a nostril and the nasal cavity into the pharynx. The optic fibers in the fiberscope are used both to illuminate the larynx and to transmit an image back to an optical sensor. Fiberscopes allow for the observation of vocal fold vibrations during running speech. Solid endoscopes, illustrated in Fig. 4.1 (right hand side), constitute of a rigid tube and a mirror inserted in the mouth and are more obstructive regarding the voicing process, but they offer a brighter image and thus a better resolution than fiberscopes (Kiritani et al., 1990).

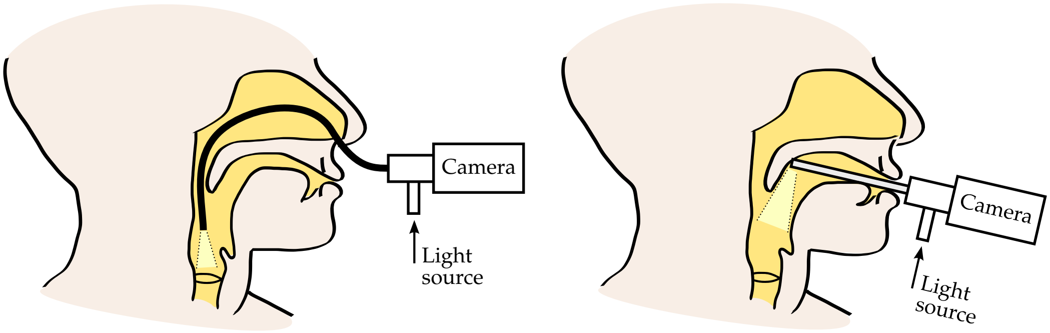


Figure 4.1: A schematic illustration of a fiberscope (on the left) and a solid endoscope (on the right).

In clinical work, laryngoscopes are usually utilized in conjunction with stroboscopic imaging. In stroboscopic imaging, the light source is flashed at a frequency close to the vocal fold fundamental frequency. Due to the flashing, the apparent motion of the vocal folds is slowed down to the difference frequency of the light source and vocal folds. This allows for easy and rapid inspection of the vibration patterns. The major drawback is that only periodic motion is captured in stroboscopy. Furthermore, stroboscopic imaging cannot be used in simultaneous recording of audio or other glottographic signals, as it is not a real-time, continuous signal.

To acquire continuous visual data on single vocal fold oscillations, high speed photography has traditionally been used (e.g. Hirano, 1981). In high speed photography, a film camera capable of exposing frames at speeds of more than 2000 frames per second is used to obtain accurate information on the vocal fold vibrations. However, since high-speed filming is expensive and the data processing is very time-consuming, use of film cameras has been replaced by digital imaging. Instead of exposing photographic film, a complementary metal-oxide semiconductor (CMOS) sensor is used to capture the images, which are then transferred to a computer for storage (Eysholdt et al., 1996). This also allows for simultaneous capture of audio and video, and even capture of colours using contemporary equipment.

Videokymography (VKG) is a special low-cost alternative for high-speed imaging. It is especially suited for analysis of vocal fold vibratory cycles. VKG uses a special charge coupled device (CCD) video camera, which can operate in two different modes. In the standard mode the apparatus works as a regular commercial video camera. In the high-speed mode, the camera delivers images from a single scan line of the whole video field at a rate of 7812.5 line images per second (Švec and Schutte, 1996; Švec et al., 2000), yielding images similar to the one in Figure 4.2. The main advantage of VKG is that the equipment required is an ordinary video camera with slight modifications, so it is significantly less expensive than proper high-speed digital imaging apparatuses. Also, the amount of data stored and processed is small when compared to high-speed imaging. On the other hand, a full image is not obtained, the measurement position has to be pre-selected, and movement of the larynx may make the recording position inaccurate. However, videokymography has been employed in basic voice research as well as in clinical practice (Švec et al., 2000).

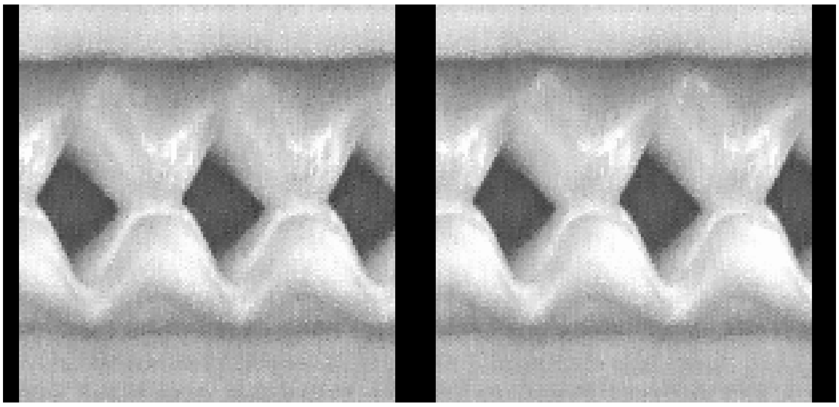


Figure 4.2: A sequence of line images obtained by videokymography. Each vertical pixel column represents a separate frame. The black blocks are the video sync periods, during which no image is obtained. Adapted with permission of the authors by Alku et al. (2000).

Photoglottography (PGG), or glottal transillumination, is another technique for acquiring an estimate of the glottal area during phonation (e.g. Hirano, 1981). In PGG, a photoelectric sensor is placed on the neck below the cricoid cartilage. The vocal folds are then brightly lit using a fiberscope. The photoelectric sensor then records the light emitted through the glottis and the neck tissue and, therefore, also the brightness variation caused by the vocal fold vibrations. The sensor and the light source may be exchanged, but the effect remains the same. The photoglottographic waveform has been found to closely correlate with the glottal area measured from high-speed image frames (Hanson et al., 1995).

Magnetic resonance imaging (MRI) is a non-invasive method using the nuclear magnetic resonance to examine features of the structure of an object. It can provide images of the structure of living tissue by using a strong magnetic field and radio frequency pulses to manipulate the spin alignment of hydrogen atoms. While the technique is not sophisticated enough to yield real-time data on vocal fold vibrations, it can be used to record vocal tract configurations, even during continuous speech (Engwall, 1999, 2004).

4.2 Electroglottography

Electroglottography (EGG) is a method for inspecting the vocal fold functioning by measuring the time-varying electrical impedance between the folds. In EGG, two electrodes, having an area typically of a few cm^2, are placed on both sides of the thyroid cartilage, and an electrical current of less than 10 mA is channeled through the electrodes. As the conductivity of the vocal fold tissue is much larger than that of the air within the laryngeal cavity and the glottis, the impedance between the electrodes varies in step with the vocal fold vibration. However, since even small direct current and low-frequency electrical currents can cause involuntary muscle contractions and great discomfort, an alternating current with a frequency of between 300 kHz and a few megahertz are typically used (Baken, 1992).

The EGG signal, since it represents the impedance between the electrodes, is relates well to the closed phase features of the glottal flow. On the other hand, the open phase is not so well detailed in the EGG signal, a sample pulse of which is shown in Fig. 4.3. The gradual increase and decrease of the contact area due to the mucosal wave and the zipper-like opening of the vocal folds is apparent in the figure, while the open phase is essentially a constant line with a slight decreasing slope due to a high-pass filter which is used to remove the large low-frequency fluctuation of the signal.

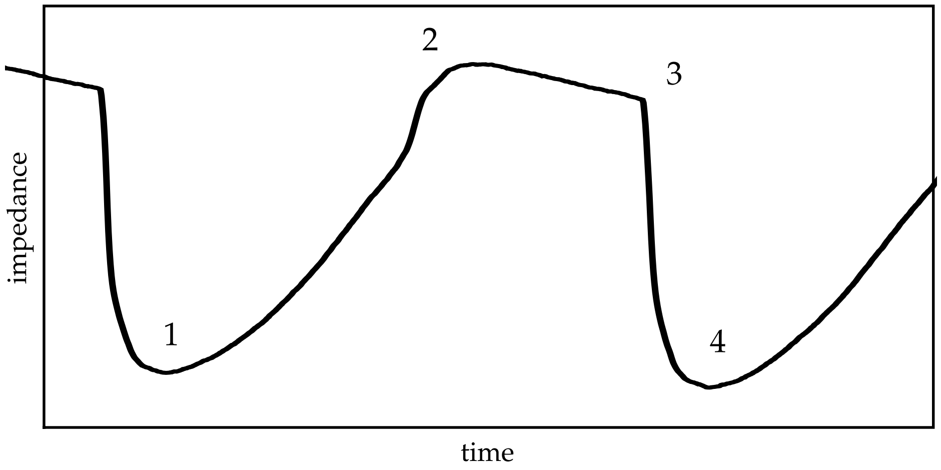


Figure 4.3: An example of an EGG signal of a male speaker in modal phonation. Point 1 indicates complete glottal closure, after which the contact area gradually decreases due to the mucosal wave and the zipper-like opening of the glottis, until at point 2 the glottis is completely open. It remains open until point 3, at which point the folds begin to close again towards point 4.

Due to its relative simplicity, EGG is a commonly used technique both in voice research and clinical work. Its qualitative assessment may reveal several forms of pathological function of voice, such as various disorders or abnormalitites of vocal fold tension. For example, different oedemas or neurological pathologies leave their mark on the EGG signal (Kitzing, 1990). Traditionally, the acquisition of the fundamental frequency (f_0) has been an important application of the EGG waveform, although modern f_0 detection algorithms have decreased the importance of this use.

Use of EGG also has several known problems, the most significant of which are the difficulties in obtaining good waveforms in female and child subjects (Colton and Conture, 1990). Furthermore, soft voice and breathy phonation cause problems to EGG, as the limited vocal fold vibration may not provide for proper contact of the vocal folds. The EGG waveform quality is also sensitive to the electrode placement and contact, which is further emphasized by laryngeal movements and neck muscle contraction during speech.

4.3 Inverse filtering

The analysis of the airflow out of the mouth and nose or even that of a microphone recording is an attractive method due to the relative simplicity and non-invasiveness of the process. Indeed, various such methods have been developed, almost all of which employ the source-filter theory of speech production, described in section 3.2, as the basis of the processing (Walker and Murphy, 2007).

The separated speech production model, illustrated in Fig. 3.3, may be expressed in the z-plane as

$$S(z) = G(z)V(z)R(z), \tag{4.1}$$

where $S(z)$ is the speech pressure waveform, $G(z)$ is the glottal flow waveform, $V(z)$ is the vocal tract filter, and $R(z)$ is the lip radiation filter. The lip radiation can be modelled accurately at low frequencies by a fixed differentiator (Flanagan, 1972). Therefore,

$R(z) = 1 - \rho z^{-1}$, where ρ is the lip radiation coefficient, whose value is close to 1. The differentiated glottal flow may be defined as the effective driving function: $Q(z) = G(z)R(z)$. Glottal inverse filtering then requires solving the following equation:

$$Q(z) = \frac{S(z)}{V(z)}. \tag{4.2}$$

Since $V(z)$ is also unknown, it has to be solved for simultaneously with $Q(z)$, which is a blind deconvolution problem. Various methods have been developed for partitioning $S(z)$ to the vocal tract and glottal flow components $V(z)$ and $Q(z)$.

Miller (1959) performed the first published inverse filtering experiments. Even though the source-filter theory of the speech production was first published by Fant (1960), Miller implicitly assumed a separated speech production model in his work. He constructed an analogue inverse network, compensating for the first two formants by utilizing prior knowledge of the formant locations and bandwidths and manually adjusting the inverse network accordingly. The method was cumbersome and was polluted by ripple of the unfiltered higher formants, but the resulting glottal flow estimates portrayed similar features as the ones obtained with contemporary methods.

Miller's method of explicit formant placement, dubbed manual inverse filtering, has been widely used both in research and in clinical work, although the recent manual inverse filtering implementations use computer software instead of analogue filter networks (e.g. Granqvist et al., 2003). Software implementations of manual inverse filtering usually allow for an arbitrary number of formants to be compensated, and since the operator's expertese is the only limiting factor in the formant placement, it can be used to effectively estimate the glottal flow even from samples otherwise deemed difficult. On the other hand, manual inverse filtering is time-consuming due to its laborious nature, and the results may be regarded more subjective than in other methods due to the reliance on the operator's judgment in the formant placement.

A popular approach to the inverse filtering problem is to utilize the closed phase of the glottal flow. If the glottis closes completely during each cycle, during that time the speech waveform must simply be a decaying oscillation affected only by the vocal tract resonances (Wong et al., 1979). Extracting the vocal tract filter $V(z)$ during the closed phase should be a simple matter, and closed phase inverse filtering has been widely used to estimate the glottal flow (e.g. Strube, 1974; Wong et al., 1979; Mataušek and Batalov, 1980; Ananthapadmanabha and Fant, 1982; Plumpe et al., 1999; Akande and Murphy, 2005). The fundamental problem of these methods is the implicit requirement of the existence of a clear closed phase. Therefore, the method is not suited to inverse filtering of very breathy voice, in which the glottal closure is only partial or non-existent.

In non-nasalized vowels, the vocal tract can be modelled with a piecewise continuous tube, the frequency response of which can be approximated with an all-pole filter structure. There exist effective methods for extracting the all-pole filter coefficients, such as the linear prediction (LP) or the digital all-pole (DAP) modelling. Such methods have been applied to the blind deconvolution problem in inverse filtering (e.g. Allen and Curtis, 1974; Milenkovic, 1986; Alku, 1992).

A popular method for utilizing *a priori* knowledge of the glottal pulse shapes is to use glottal flow models such as those described in section 3.2 in the inverse filtering process. Several such approaches have been proposed (e.g. Krishnamurthy, 1992; Kasuya et al., 1999; Fröhlich et al., 2001; Arroabarren and Carlosena, 2003; Bozkurt et al., 2005).

Several external aids have been developed for the inverse filtering process. Rothenberg (1973) developed a pneumotachograph, or Rothenberg mask, to estimate the airflow out of the mouth, thus avoiding the effects of the lip radiation effect. This allows for measurement of absolute flow values, which is not possible in inverse filtering of the speech pressure signal. Rothenberg used manual inverse filtering to remove the formants from the flow signal, but in principle most inverse filtering methods could easily be adapted for use with a Rothenberg mask. Its main limitation is the small bandwidth of only 1.6 kHz (Hertegård and Gauffin, 1992), which causes the glottal flow pulses to become artificially rounded and devoid of features. Furthermore, in many recording tasks the use of the mask is restrictive and limits expression.

The Sondhi tube is another tool for glottal flow measurement. Sondhi (1975) noted that by speaking into a reflectionless uniform tube the vocal tract resonances are considerably reduced, and in the case of a neutral vowel, even almost entirely eliminated. Thus, the glottal waveform can be directly recorded using a microphone inserted in the tube wall. The use of the Sondhi tube restricts the uttered sounds only to the neutral vowel, which makes it unsuitable for most speech tasks.

There also exist methods in which signal paths other than the microphone recordings are used in the separation of the voice source and the vocal tract. Especially the use of EGG signals has been found beneficial together with the closed-phase inverse filtering methods, since EGG is able to indicate the glottal closure and opening instants much more precisely than the microphone recording (e.g. Veeneman. and BeMent, 1984; Krishnamurthy and Childers, 1986).

Evaluation of inverse-filtering methods has been a long-standing problem. Direct measurement of the glottal flow is not practically possible, so direct comparisons between the flow estimates and the actual flow cannot be made. Several other techniques, however, have been used to assess the correctness of inverse filtering methods. Glottal waveform estimates have been compared to other simultaneously acquired signals such as EGG, VKG, or high-speed imaging (e.g. Baer et al., 1983; Childers et al., 1984; Granqvist et al., 2003). While these techniques assess the vocal fold behaviour, they do not measure the glottal airflow, but just some related quantities. No one-to-one mapping exists between the glottal flow and their values, and therefore information given by them is always limited.

5 Voice source parameterization

Analytic study of the voice source requires not only use of some voice analysis techniques, as described in the previous section, but also parameterization of the data acquired in the analysis. Parameterization implies quantifying the obtained signals with properly selected numeral values (Alku, 2003). In parameterization, some significant features of the original signals are represented in a compressed numerical form. This compact representation of the signal features can then be used in quantitative analysis and study of voice production.

5.1 Absolute measurements of glottal properties

The most obvious way to parameterize glottal behaviour would be to measure different properties of the vocal folds or the glottis during the vibratory process. Laryngoscopy yields an optical image of the vocal folds from which the measurement of the glottal area or the vocal fold length is comparatively straightforward (Childers et al., 1983; Eysholdt et al., 1996). This process is also illustrated in Figure 5.1. Unfortunately, the exact scale of the image is usually not defined due to the unknown distance between the optical lens and the vocal folds, and therefore only relative measures are normally acquired. Techniques have been developed to perform distance calibration using laser triangulation or stereoscopic imaging, yielding absolute measures of the vocal folds (Hertegård et al., 2003; Wittenberg et al., 2000).

When the Rothenberg mask is used, absolute measurements of the glottal airflow can be made using waveforms estimated by inverse filtering. Parameters typically used include the peak, minimum, mean, and peak-to-peak flow (e.g. Holmberg et al., 1988).

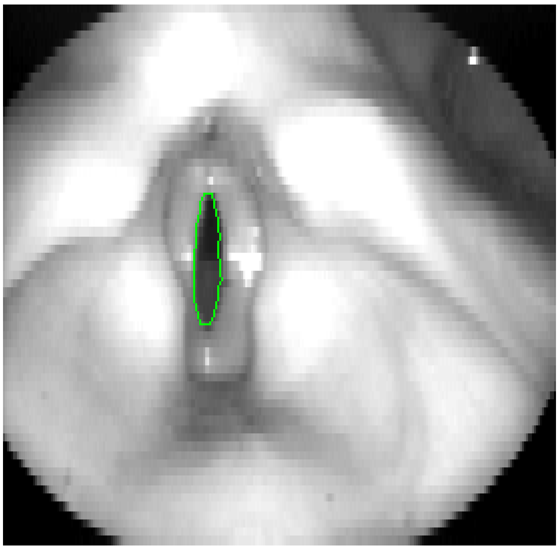


Figure 5.1: A laryngoscopic measurement of the glottal area with the perimeter of the glottis traced in the High-speed Tool Box software (Larsson et al., 2000). Image courtesy of Hannu Pulakka.

However, most inverse filtering experiments use microphone signals due to their less disruptive nature, and in those experiments absolute flow amplitude measurements are not possible.

5.2 Time-based parameters

Timcke et al. (1958) were the first to express the temporal properties of the glottal fold vibration, as observed from ultra-high-speed motion pictures, in a parameterized manner. They defined the open quotient (OQ) and speed quotient (SQ) parameters as

$$\mathrm{OQ} = \frac{T_{\mathrm{op}}}{T} \tag{5.1}$$

and

$$\mathrm{SQ} = \frac{T_{\mathrm{o}}}{T_{\mathrm{c}}}, \tag{5.2}$$

where T_{op} is the time when the glottis is open and T is the total period duration, T_{o} is the duration of the opening period, and T_{c} the duration of the closing period. Furthermore, $T_{\mathrm{op}} = T_{\mathrm{o}} + T_{\mathrm{c}}$. OQ and SQ have since been used repeatedly in the parameterization of the glottal area obtained by high-speed imaging, the glottal flow volume velocity waveform obtained by inverse filtering, and in the case of OQ, EGG waveforms. OQ has been found to be negatively correlated with the intensity and the pressedness of speech, while SQ increases as the intensity increases (Holmberg et al., 1988; Childers and Lee, 1991). However, parameters relying on the determination of the maximal opening instant such as SQ can be problematic to determine when using EGG since the glottal opening may be difficult to detect. Since the various signals provide complementary but differing information regarding the voice production, direct comparisons of the parameters computed from the different signals may be difficult. For example, the area function waveforms tend to have more symmetrical opening and closing phases than the flow waveforms, in which the closing phase is usually considerably shorter. These differences directly affect the SQ. Rows (c) and (d) in Figure 5.2 illustrate the measurements of the open and speed quotients and the effects of different values of the parameters on the parameter shape.

Monsen and Engebretson (1977) introduced the closing quotient (ClQ), defined as

$$\mathrm{ClQ} = \frac{T_{\mathrm{c}}}{T}, \tag{5.3}$$

and illustrated in Figure 5.2 (e). Its use is well-founded, since the majority of the voicing energy stems from the abrupt closure of the glottis occurring at the end of the closing phase. Therefore, the properties of the closing phase most directly affect the voice quality. The closing quotient has been found to decrease when the vocal intensity and the pressedness increase (e.g. Holmberg et al., 1988).

A common problem for all time-based parameters is that the exact locations of the events such as the glottal opening or closing instant are often vague and subject to in-

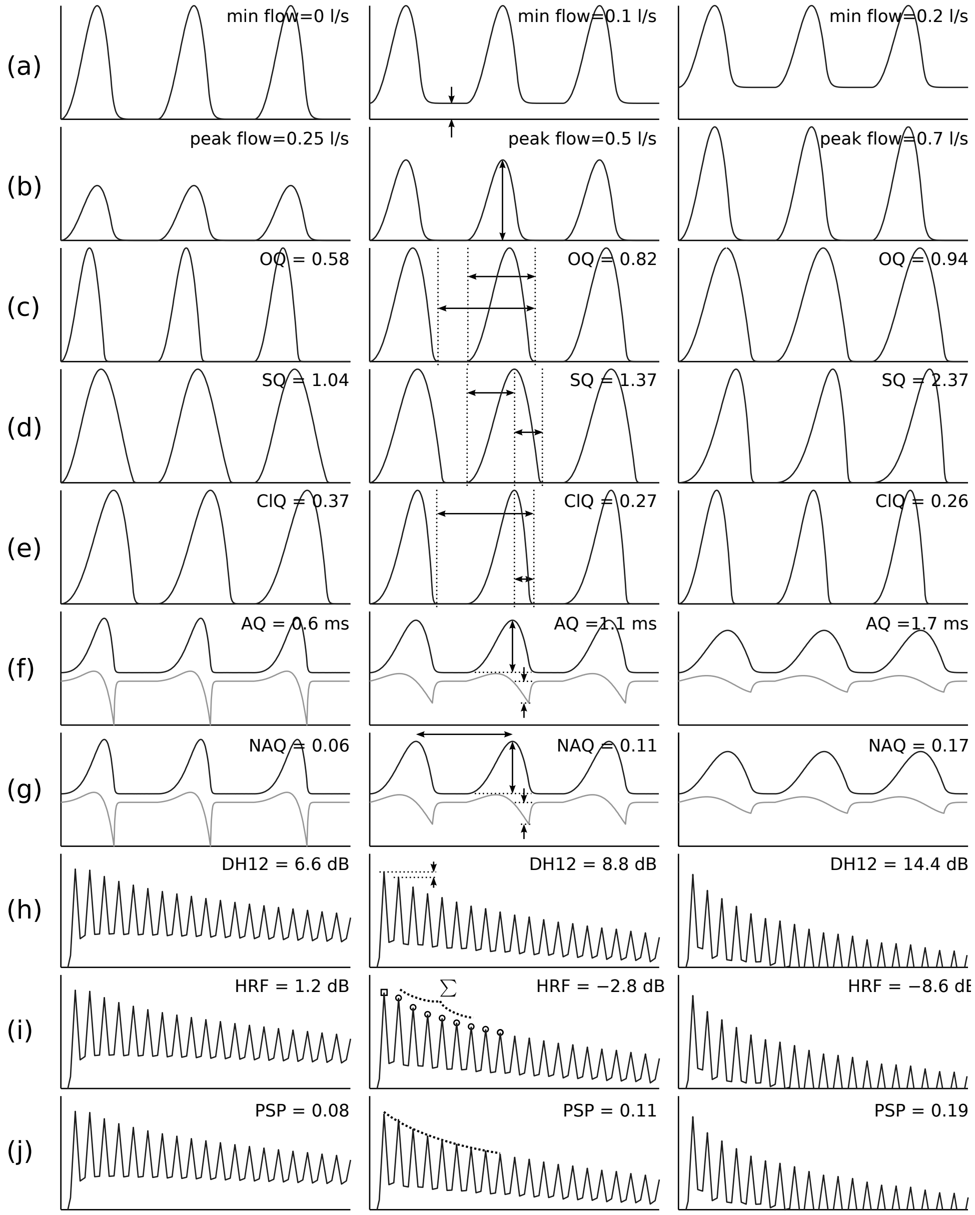


Figure 5.2: Schematic representations of different common voice source parameters. Figures representing three different values are shown for each parameter. Rows (a)–(g) represent parameters of the time-based glottal waveform, while (h)–(j) illustrate parameters based on the magnitude spectrum of the glottal flow. The lower curves in figures (f) and (g) represent the flow derivative.

terpretation due to either the smoothness of the pulse or residual formant ripple stemming from incomplete cancellation of the formants (Dromey et al., 1992; Holmberg et al., 1988). This reduces the precision and robustness of these parameters. One way to avoid the problem is to combine amplitude-based time instants so that they express properties related to the time domain of the signals. One such parameter is the amplitude quotient (AQ), defined as

$$\mathrm{AQ} = \frac{A_{\mathrm{ac}}}{d_{\mathrm{min}}}, \tag{5.4}$$

where A_{ac} is the flow peak-to-peak amplitude (the difference between the maximum and the minimum value within one period) and d_{min} is the minimum value of the flow derivative (Alku and Vilkman, 1996a). Sample values of AQ are shown in Figure 5.2 (f). The measurements are straightforward to obtain, and the absolute scale of the glottal flow pulses is not required to be known.

While the AQ parameter has been shown to correlate with the phonation type, in some measurements its strong dependence on the f_0 of the signal may be found problematic. Therefore, a simple f_0 normalization of the parameter has been proposed, resulting in the normalized amplitude quotient (NAQ) (Alku et al., 2002), shown in Figure 5.2 (g). The NAQ is defined as

$$\mathrm{NAQ} = \frac{\mathrm{AQ}}{T}. \tag{5.5}$$

Fant (1995) has also arrived at essentially the same parameter (R_d) using a set of transformed LF-model parameters. The NAQ parameter has been shown to correlate well with the expression of the phonation type in intensity changes but to be more robust than the ClQ parameter (Bäckström et al., 2002).

A common method of glottal flow parameterization is to fit a synthetic glottal flow model waveform over the inverse filtered glottal flow waveform. The LF-model has traditionally been the most popular model in fitting tasks (e.g. Strik et al., 1993). However, no technical reason prevents from using any other glottal flow model instead (e.g.

Veldhuis, 1998; Doval et al., 2006). When an optimal fit is acquired, the model parameters are then used to describe the original glottal flow waveform. The model-based parameterization methods are especially useful when the vowels need to be resynthesized using the acquired parameters, but otherwise the forceful fitting of the waveform to a synthetic model may cause data to be lost unnecessarily.

5.3 Spectral parameters

Small imperfections of the inverse filtering process may lead to gross disfiguration of the time-domain glottal flow waveform. However, the power spectrum of the waveform may still be largely unchanged. Several parameters have been proposed to facilitate parameterization of the spectra of the glottal flow pulses. $\Delta\mathrm{H}_{12}$, or H1-H2, is the difference of the first two harmonics on the decibel scale (Titze and Sundberg, 1992). The harmonic richness factor (HRF), defined as

$$\mathrm{HRF} = \frac{\sum_{k\geq 2} H_k}{H_1}, \tag{5.6}$$

where H_k is the kth harmonic, is closely related to $\Delta\mathrm{H}_{12}$, but tries to approximate the spectral energy distribution from more than one higher harmonic (Childers and Lee, 1991). Attempts to further improve spectral parameterization include the parabolic spectrum parameter (PSP), in which a second-order polynomial is fitted to the harmonics to gain an estimate of the spectral slope (Alku et al., 1997). The $\Delta\mathrm{H}_{12}$, HRF, and PSP parameters are illustrated in Figure 5.2.

6 Voice quality in vocal expression of emotion

Voice quality, defined as the characteristic auditory colouring of voice as described in Section 1, is a major component of the non-verbal information constituting the so-called second channel of speech. Its importance is especially pronounced in vocal expression of emotions, which effectively communicates important additional information on the speaker's stances, attitudes, and intentions.

6.1 Definition and properties of voice quality

Voice quality is a concept encompassing the characteristic auditory colouring of a speaker's voice related to their identity, personality, health, and emotional state (Story et al., 2001). It is a broad term and has been further defined as "those characteristics which are present more or less all the time that a person is talking" (Abercombie, 1967, p. 91). These characteristics are created by all subprocesses of speech production, i.e. the respiratory, laryngeal, and articulatory systems.

Laver (1980) created an influental phonetic classification system for describing different voice qualities. His model describes different laryngeal and supralaryngeal settings of the voice production system and uses those physiological changes as a method for describing the voice quality. Supralaryngeal settings are divided into longitudinal, latitudinal, and velopharyngeal settings. Laryngeal, or phonatory, settings include modal voice, falsetto, whisper, creak, harshness, and breathiness, and several compound phonation types. Furthermore, Laver defined general tension settings, i.e. tense and lax voice, which affect the overall muscular tension throughout the vocal system.

In speech research, a narrower definition of voice quality is often preferred in which the voice quality is considered to stem only from laryngeal and respiratory features, thus ignoring the effects of the vocal tract. This narrower definition is especially suitable in the studies of the voice source, where the voice source signal can be considered to incorporate all voice quality features in itself. Even when the narrower definition is used, Laver's settings can be applied to describe the laryngeal voice quality (e.g. Gobl and Ní Chasaide, 1992). However, simplified descriptions are often used, encompassing only a single axis of voice quality variation, such as breathy, normal, and pressed phonation (e.g. Alku and Vilkman, 1996b), or modal voice, vocal fry, and breathy voice (Childers and Ahn, 1995). Such descriptions are useful as the changes can be quantified using a single voice source parameter.

Voice quality is an omnipresent, pervasive property of speech. Together with fundamental frequency and intensity changes, it acts as a "second channel" of speech, conveying extralinguistic information regarding moods, attitudes, emotions, health, and physical properties of the speaker. Although in regular speech the voice quality changes are automatic and largely subconscious, they can also be consciously controlled, at least to a degree. For example, voice quality can convey culturally specific affect content, or a mismatch between the lexical meaning, as in expression of humour, sarcasm, or irony (Gobl, 2003).

Laryngeal voice quality changes have been connected to variations in the voice source parameters in several studies. Klatt and Klatt (1990) found the breathy voice quality to be signalled by a number of acoustic cues, such as a constant flow component, increased open quotient, amplitude of the first harmonic, and first-formant bandwidth, as well as reduced amplitudes of higher harmonics and a less distinct first-formant peak. Childers and Lee (1991) found the HRF to be high in vocal fry, medium in modal voice, and low in falsetto and breathy voice. Similar relations were observed for OQ, SQ, and the LF-model parameter t_a, breathy and falsetto exhibiting rounded pulse shapes and vocal fry very sharp pulses. The results were further supported by the study of the LF-model parameters by Gobl and Ní Chasaide (1992) and Childers and Ahn (1995) as well as of the NAQ parameter by Alku et al. (2002) and Bäckström et al. (2002).

6.2 Vocal expression of emotions

Emotions are processes of events, affecting several psychophysiological components of an organism: physiological arousal, motor expression, and subjective feeling (Scherer, 2000). They represent an organized, highly structured reaction to an event that is relevant to the needs, goals, or survival of the organism (Watson, 2000). As emotions induce a physiological response that can be observed by others, they also have inherent communicative properties. Emotions present themselves in general appearance, such as facial expressions and body postures, as well as in vocal expression. The expressions can be interpreted with great accuracy by other individuals.

Different models for classifying emotions have been proposed. For example, several popular theories list basic emotions, most of which include joy, interest, surprise, fear, anger, sadness, and disgust as a core set, which should be considered more primitive and universal than others (Ekman and Davidson, 1994). Structural models regard emotions as comprising of different component processes, the precise combination of which define the actual emotions (e.g. Scherer, 2003). A popular emotion classification method is to set the emotions along a few abstract dimensions, the most popular of which is the activation-valence (or activation-evaluation) space (Schlosberg, 1954). In this scheme, valence rates whether the emotion is a positive or a negative one and activation describes the emotion on a active-passive scale. This method is often practical in describing the differences between the emotions, even though it is unable to separate some key emotions such as anger and fear properly.

In speech, emotions are elicited at suprasegmental, segmental, as well as intrasegmental levels (Murray and Arnott, 1993). While all these levels contain both verbal and vocal information, the emotional expression in the vocal, “implicit” channel of speech production can render the verbal information redundant, qualify it further, or contradict it. At the suprasegmental and segmental levels, emotions are expressed by the fundamental frequency, intensity, and segment duration patterns. Even though these are important factors in vocal expression of emotions, studies on emotional speech synthesis have indicated that the presence of intrasegmental voice quality adjustments improve the quality of, or are even by themselves sufficient to elicit emotional expression (Burkhardt, 2000; Gobl and Ní Chasaide, 2003).

Traditionally, the research on vocal expression of emotions has mostly utilized parameters related to f_0, intensity, and duration changes (e.g. Murray and Arnott, 1993; Banse and Scherer, 1996; Cowie et al., 2001). While the significance of voice quality changes has been acknowledged, their use in emotion analysis was limited due to cited methodological issues (Scherer, 1986). Relative frequency band energy contents, correlating to spectral tilt variations of different voice qualities, have been used as factors in some studies (e.g. Banse and Scherer, 1996). However, studies regarding inverse filtering of emotional speech have indicated that even voice quality changes alone evoke significant emotional colourings in otherwise neutral utterances, although there is no one-to-one mapping between different emotions and voice qualities (Gobl and Ní Chasaide, 2003). Instead, the voice qualities are used in conjunction with other speech features to express specific emotions.

Murray and Arnott (1993) presented a summary of the effects of vocal expression of different emotions, given in Table 6.1. Similar effects have also been reported by Banse and Scherer (1996) and Cowie et al. (2001). While some emotions, such as anger and sadness or happiness and disgust, show strong differences in the basic acoustic dimensions, the table indicates very similar profiles for anger, happiness, and fear regarding the speech rate, pitch average and range, and intensity variables. However, when laryngeal voice quality changes are inspected, marked differences become apparent (Ní Chasaide and Gobl, 2004). For example, anger is related with a very tense voice quality, while sadness is expressed with a lax-creaky voice, and fear with a breathy or whispery voice. These differences allow for a much better separation of emotions when voice quality changes are taken into account.

	Anger	Happiness	Sadness	Fear	Disgust
Speech rate	slightly faster	faster or slower	slightly slower	much faster	very much slower
Pitch average	very much higher	much higher	slightly lower	very much higher	very much lower
Pitch range	much wider	much wider	slightly narrower	much wider	slightly wider
Intensity	higher	higher	lower	normal	lower
Voice quality	breathy, chest tone	breathy, blaring	resonant	irregular voicing	grumbled, chest tone
Pitch changes	abrupt, on stressed syllables	smooth, upward inflections	downward inflections	normal	wide, downward terminal inflections
Articulation	tense	normal	slurring	precise	normal

Table 6.1: Summary of effect of emotions on speech (after Murray and Arnott, 1993).

7 Summary of publications

This thesis comprises seven publications of which five were published in international reviewed journals and two in reviewed conference proceedings. The articles are divided in two groups. Articles I and II address voice quality characteristics in different contexts, while Articles III–VII discuss voice source analysis and parameterization methods.

Publication I: ”Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient”

In this paper, emotional expression in short vowels segments of continuous speech were analyzed. The main goal of the study was to establish quantitative data on the voice source behaviour in the expression of emotions. As material, acted emotional portrayals by nine professional actors (five males, four females) were used. The five emotions used—neutral, sadness, joy, anger, and tenderness—were chosen so that they are well separated in the activation-valence space, and that the acoustic and perceptual differences could be presumed to be as large as possible. Recited passages of Finnish prose were recorded, inverse filtered, and parameterized using the NAQ parameter. Statistical analysis using linear regression (ANOVA) indicated that both emotion and speaker sex had a significant effect on the NAQ value. Post-hoc analysis indicated that there are significant differences between most emotion pairs. It was found that anger was expressed with the most pressed voice (indicated by the smallest NAQ value), followed by neutral, joy, sadness, and tenderness. The pressed/breathy voice quality, indicated by the NAQ values, were found to correlate with the activation dimension of the emotions. Furthermore, it was found that the female speakers expressed considerably wider variations both within and between emotions.

Publication II: ”Laryngeal voice quality changes in expression of prominence in continuous speech”

The second study concerned voice quality changes in expression of prominence. Here, prominence (stress) reflects to the contrasting prosodic properties of nonverbal expression, which are used to emphasize some linguistic elements such as answers to questions or the main topic of conversation. The hypothesis of the work was that prominence in speech is expressed with a more pressed voice quality so that there would be significant voice quality differences between stressed and unstressed syllables and words. This was tested by recording the speech of 11 speakers. The text recited by speakers was chosen so that suitable vowels in different lexical positions could be picked out. The vowels were inverse filtered and parameterized with TKK Aparat (see Publication III) using NAQ and AQ parameters both of which are negatively correlated with pressedness of speech. The study did not involve measurements of subglottal pressure or intensity. Contradictory to the initial research hypothesis, the results indicate that stressed vowels are expressed with a breathier voice quality than unstressed vowels. The result was explained by features of speech production physiology: in continuous speech, the rapid voice quality changes required for expression of stress cannot be performed due to slow adjustments of sub-glottal pressure. Therefore, only f_0 is increased, but as the total speech production power remains the same, the energy available for each glottal cycle decreases leading to more rounded glottal flow waveforms and a breathier voice quality. The results suggest that studies performed on sustained vowels or artificial voicing tasks might not be always applicable to continuous speech.

Publication III: ”TKK Aparat: An Environment for Voice Inverse

Filtering and Parameterization”

This article describes TKK Aparat, a voice source inverse filtering and parameterization software toolkit. As relatively few voice inverse filtering packages have been published, a special software for this purpose was constructed. A simplified version of iterative adaptive inverse filtering (IAIF) called direct inverse filtering (DIF) is proposed, and both inverse filtering algorithms are described. The method and results of evaluating the inverse filtering algorithms by comparing a large number of inverse filtering parameters is described. The implemented parameters are also discussed together with the detailed description of the methods to acquire the parameterization time instants. The essential user interface elements including the inverse filtering, parameterization, and visualization features are introduced. Usability testing was also performed to reveal any usability issues and improve the interface. Finally, different projects in which TKK Aparat has already been used are described. The software package has already shown to be a useful tool, and has been adopted by several speech research groups.

Publication IV: ”Comparison of Multiple Voice Source Parameters in Different Phonation Types”

Article IV is a comparison of different voice source parameters. Although numerous glottal flow parameters have been proposed, few quantitative comparisons of the parameters have been made. In this paper, stationary utterances of all eight Finnish vowels in three different phonations (breathy, normal, and pressed) where acquired from 11 speakers. The data was inverse filtered and parameterized with TKK Aparat using all 21 parameters supported by the software. Statistical analysis of the parameters was performed by computing a linear regression model for every parameter. Then, the proportions of variation explained were calculated to find out which parameters were able to best reflect the phonation changes. Furthermore, correlation matrices were computed to discover which parameters are well related to each other regarding the phonation type. Parameters focusing on the glottal closing phase such as NAQ, AQ, and ClQ were able to express the phonation type best. Cross-correlation matrices indicated that parameters tend to correlate with other members of the same parameter group. For example, different closing phase parameters were well correlated, as were the different OQ parameters. The results are useful in assisting in the selection of suitable voice source parameters when the phonation changes are to be quantified. Furthermore, the results assist in making comparisons of the results of existing papers in which different voice source parameters have been used.

Publication V: ”Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production”

In this study, the goal was to test a novel method to estimate the performance of an inverse filtering technique using physical modelling of voice production. Speech pressure signals are generated using a combined two-mass model of the vocal folds and a wave reflection (digital waveguide) model of the trachea and the vocal tract. The mass-spring model was aerodynamically coupled with the digital waveguide to provide a realistic interaction between the two elements. The speech production model was used to produce four different vowels, each with ten different values of the fundamental frequency. Since both the generated vowels and their respective voice source signals were acquired from the synthetic model, they were able to act as a reference for the inverse filtering evaluation. The inverse filtering method tested in this study was the iterative adaptive inverse filtering (IAIF), which was used to inverse filter the synthetic vowels to create glottal flow estimates. Both the synthetic and estimated glottal flows were then parameterized using the normalized amplitude quotient and the harmonic level difference (H1-H2). The results indicated that the error introduced by inverse filtering was, in general, small for both parameters. The effect of the distortion caused by inverse filtering on the parameter values was clearly smaller than the change in the corresponding parameters when the phonation type was altered. The error caused by inverse filtering was largest for high-pitched vowels with low first formants, as expected. The study was able to show that the errors induced by inverse filtering are sufficiently small that the proposed inverse filtering technique is able to measure the voice source dynamics with satisfactory accuracy.

Publication VI: ”Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types”

In this article, two common inverse filtering methods are compared to observe whether the inverse filtering method affects parameterization of the voice source. Speech pressure waveforms of six female and seven male speakers producing sustained /a/ vowels in breathy, normal, and pressed phonations were recorded. The recordings were inverse filtered by two different methods: manual inverse filtering, as implemented in the DeCap software, and the semiautomatic iterative adaptive inverse filtering (IAIF) method. Both of the methods were used by three different speech research professionals. The closing phase characteristics of the estimated glottal flow waveforms were parameterized using two time-based parameters, ClQ and NAQ. Statistical analyses were then performed on the acquired parameters. Since normality tests indicated that the data was not normally distributed, nonparametric tests were used. According to the statistical analyses, statistically significant differences between the inverse filtering methods were present. However, the correlation of the results using the two methods were very high and the discrepancies, in general, reasonably small. Therefore, the result of this study was found to be encouraging in showing that automatic inverse filtering can be developed in the future to meet the needs of extensive speech data analysis.

Publication VII: ”An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity”

This study presents an approach to visualizing intensity regulation in speech. The method expresses voice samples in two-dimensional space using amplitude-domain values extracted from inverse filtered glottal flow estimates. The proposed presentation expresses a time-domain measure of the glottal pulse, the amplitude quotient (AQ), as a function of the negative peak amplitude of the flow derivative (d_{peak}). Voice samples varying from very soft to very loud with a SPL range of approximately 55 dB were analyzed using the proposed method. The results indicate that when vocal intensity is increased, the speech samples first showed a rapidly decreasing trend on the graph. When intensity is further increased, the samples converged toward a horizontal line, the asymptote of the regression hyperbola. The behaviour of the AQ-d_{peak} graph indicates that the intensity regulation strategy changes from laryngeal to respiratory mechanisms as the intensity increases. The proposed analysis method makes it possible to quantify how the control mechanisms underlying intensity regulation change gradually between the laryngeal and respiratory mechanisms. The proposed presentation constitutes a concise visualization method for the intensity regulation functioning because the only information needed is the glottal flow waveform estimate inverse filtered from the acoustic speech pressure signal.

8 Conclusions

This thesis sought new ways of improving the voice source analysis methodology in voice quality studies and data of the actual voice quality changes in continuous speech. The methodological issues were addressed in five different publications, while two papers were studies on the voice quality changes.

In the course of this work, a novel software environment for voice source inverse filtering and parameterization, *TKK Aparat*, was created. This work was described in Article III. The algorithms and parameters implemented in this software were mostly already published, but the software has practical importance as a common platform for voice source studies and as a reference implementation of multiple principal algorithms in the field. The software has received a favourable response and has been already adopted by many research groups. The other publications regarding the methodology have further tested the validity of the voice source analysis methods used (Article VI) and proposed new techniques for testing them (Article V). Furthermore, guidelines for the glottal volume flow waveform parameterization have been set in Article IV, and new methods for interpretation of the glottal flow parameters are suggested in Article VII.

The results of the methodological articles should prove useful in many works concerning the study of the voice source. For example, the software created can be, and has been, applied to studies of occupational voice and laryngology. The validation studies can be repeated using other algorithms, thus providing comparable data on the similarities of these algorithms and their applicability. Article IV, regarding different voice source parameters, provides important guidelines for future research, including selection of algorithms, as well as giving information on the comparability of different glottal flow parameters. This information should prove valuable to anyone working on the speech pressedness and voice source analysis. Finally, the novel way of visualizing the function of voice production suggested in Article VII should supplement existing voice production analysis tools and therefore prove useful in the research of vocal function.

The goal of acquiring data on the voice source behaviour in voice quality changes was addressed in two studies. In the first of these publications, Article I, voice source changes in expression of emotions were studied. This study suggested a correlation between the pressedness of the voice and the activation dimension of emotional expression. In Article II, voice quality changes in expression of prominence (stress) in speech were studied. This research yielded unexpected results contradicting the initial intuitive hypothesis. A new, plausible explanation for the observed phenomenon was suggested.

As a thorough analysis of the different voice quality issues with regard to the voice source function would have extended beyond the realms of a single doctoral thesis, only two seminal issues were studied. Both of the Articles I and II should provide useful information for generation of expressive speech synthesis. Article I gives an insight on the voice quality changes present in expression of different emotions, while Article II should be useful in the generation of physiologically motivated speech intonation models. Incorporation of voice quality changes in these domains should improve the expressivity of synthetic speech.

In no way is this research complete. First, the practical limitation of only studying the changes on the breathy–pressed axis of the voice quality rejects other important voice quality features. Methods for their analysis have been developed, and they should be incorporated in future studies to acquire a more thorough view of the voice quality changes in different situations. Second, although this work improves on the unfortunately common tradition of the field of using extremely limited amounts of data, even larger data sets could have been used. Therefore, the results of this work should, in most cases, not be considered confirmatory, but exploratory in nature. To acquire authoritative results, larger data sets should be analysed. Hence, fully automatic voice inverse filtering and voice quality measurement methods need to be developed to facilitate the analysis of large data quantities.

The author believes that this work, on its own part, will help in facing the methodological challenges which so far have been perceived as hindering the analysis of the voice qualities in many fields of research. Yet, the inherently multidisciplinary nature of the voice source analysis renders it a challenging and complex domain.

References

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

O. Akande and P. Murphy. Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Commun*, 46:15–36, 2005.

F. Alipour. A finite-element model of vocal-fold vibration. *J Acoust Soc Am*, 108(6): 3003–3012, 2000.

F. Alipour and I. R. Titze. Simulation of particle trajectories of vocal fold tissue. In *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, pages 183–190, Denver, Colorado, 1985.

P. Alku. Parameterisation methods of the glottal flow estimated by inverse filtering. In *Proc ISCA Workshop on Voice Quality: Functions, analysis and synthesis (VOQUAL03)*, pages 81–87, Geneva, Switzerland, 2003.

P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Commun*, 11:109–118, 1992.

P. Alku and E. Vilkman. Amplitude domain quotient of the glottal volume velocity waveform estimated by inverse filtering. *Speech Commun*, 18:131–138, 1996a.

P. Alku and E. Vilkman. A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. *Folia Phoniatr Logo*, 48:240–254, 1996b.

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

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

P. Alku, J. G. Švec, E. Vilkman, and F. Šram. Analysis of voice production in breathy, normal and pressed phonation by comparing inverse filtering and videokymography. In B. Yuan, T. Huang, and X. Tang, editors, *Proceedings of the 6th International Conference of Spoken Language Processing (Interspeech 2000)*, pages 885–888, Beijing, China, 2000. China Military Friendship Publish.

P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *J Acoust Soc Am*, 112(2):701–710, 2002.

J. B. Allen and T. H. Curtis. Automatic extraction of glottal pulses by linear estimation. *J Acoust Soc Am*, 55(2):396, 1974.

T. V. Ananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *Speech Commun*, 1:167–184, 1982.

I. Arroabarren and A. Carlosena. Glottal spectrum based inverse filtering. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 57–60, 2003.

F. Avanzini. Simulation of vocal fold oscillation with a pseudo-one-mass physical model. *Speech Communication*, 50(2):95–108, 2008.

F. Avanzini, P. Alku, and M. Karjalainen. One-delayed-mass model for efficient synthesis of glottal flow. In *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 51–54, Aalborg, Denmark, 2001.

T. Baer, A. Löfqvist, and N. S. McGarr. Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques. *J Acoust Soc Am*, 73(4):1304–1308, 1983.

R. J. Baken. Electroglottography. *J Voice*, 6(2):98–110, 1992.

R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol*, 70(3):614–636, 1996.

E. Björkner. *Why so different? Aspects of voice characteristics in operatic and musical theatre singing*. PhD thesis, KTH, Stockholm, Sweden, 2006.

B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit. Zeros of Z-transform representation with application to source-filter separation in speech. *IEEE Sig Proc Let*, 12 (4), 2005.

T. Brancatisano, P. W. Collett, and L. A. Engel. Respiratory movements of the vocal cords. *J Appl Physiol*, 54(5):1269–1276, 1983.

F. Burkhardt. *Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren*. PhD thesis, Technische Universität Berlin, Berlin, Germany, 2000.

T. Bäckström, P. Alku, and E. Vilkman. Time-domain parametrization of the closing phase of glottal airflow waveform from voices over a large intensity range. *IEEE T Speech Audi P*, 10(3):186–192, 2002.

D. Childers and C. Ahn. Modeling the glottal volume-velocity waveform for three voice types. *J Acoust Soc Am*, 97:505–519, 1995.

D. G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *J Acoust Soc Am*, 90(5):2394–2410, 1991.

D. G. Childers, J. M. Naik, J. N. Larar, A. K. Krishnamurthy, and G. P. Moore. Electroglottography, speech, and ultra-high speed cinematography. In I. R. Titze and R. C. Scherer, editors, *Vocal Fold Physiology, Biomechanics, Acoustics and Phonatory Control*, pages 202–220, The Denver Center for the Performing Arts, Denver, Colorado, USA, 1983.

D. G. Childers, A. M. Smith, and G. P. Moore. Relationships between electroglottograph, speech, and vocal cord contact. *Folia Phoniatr*, 36(3):105–18, 1984.

R. H. Colton and E. G. Conture. Problems and pitfalls of electroglottography. *J Voice*, 4(1):10–24, 1990.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Proc Mag*, 18(1):32–80, 2001.

K. Dedouch, J. Horácek, T. Vampola, J. Svec, P. Kršek, and R. Havlık. Acoustic modal analysis of male vocal tract for czech vowels. In *Proceedings of Interaction and Feedback 2002*, pages 13–20, Prague, Czech Republic, 2002.

B. Doval, C. d'Alessandro, and N. Henrich. The spectrum of glottal flow models. *Acta Acust United Ac*, 92(6):1026–1046, 2006.

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

P. Ekman and R. J. Davidson, editors. *The nature of emotion: Fundamental questions*. Oxford University Press, New York, USA, 1994.

A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Trans Sig Process*, 39:411–423, 1991.

O. Engwall. From real-time MRI to 3D tongue movements. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004)*, volume 3, pages 1693–1696, Jeju island, Korea, 2004.

O. Engwall. Modeling of the vocal tract in three dimensions. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech'99)*, pages 113–116, Budapest, Hungary, 1999.

U. Eysholdt, M. Tigges, T. Wittenberg, and U. Pröschel. Direct evaluation of high-speed recordings of vocal fold vibrations. *Folia Phoniatr Logo*, 48(4):163–170, 1996.

G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, Netherlands, 1960.

G. Fant. *Speech Sounds and Features*. MIT Press, Cambridge, Massachusetts, USA, 1973.

G. Fant. The LF-model revisited. transformations and frequency domain analysis. *STL-QPSR*, (2–3):119–156, 1995.

G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, (4):1–13, 1985.

W. Fitch and J. Giedd. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *J Acoust Soc Am*, 106(3):1511–1522, 1999.

J. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, 1972.

J. L. Flanagan and L. L. Landgraf. Self-oscillating source for vocal-tract synthesizers. *IEEE T Audio Electroacoust*, AU-16(1):57–64, 1968.

M. Fröhlich, D. Michaelis, and H. W. Strube. SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *J Acoust Soc Am*, 110: 479–488, 2001.

C. Gobl. *The voice source in speech communication*. PhD thesis, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, 2003.

C. Gobl and A. Ní Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech Commun*, 40:189–212, 2003.

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

S. Granqvist, S. Hertegård, H. Larsson, and J. Sundberg. Simultaneous analysis of vocal fold vibration and transglottal airflow: exploring a new experimental setup. *J Voice*, 17(3):312–330, 2003.

H. Gray and W. H. Lewis. *Anatomy of the Human Body*. Lea & Febiger, Philadelphia, USA, 1918.

H. E. Gunter. A mechanical model of vocal-fold collision with high spatial and temporal resolution. *J Acoust Soc Am*, 113(2):994–1000, 2003.

A. Hannukainen, T. Lukkari, J. Malinen, and P. Palo. Vowel formants from the wave equation. *JASA Expr Lett*, 122(1):1–7, 2007.

D. G. Hanson, J. Jiang, M. D'Agostino, and G. Herzon. Clinical measurement of mucosal wave velocity using simultaneous photoglottography and laryngostroboscopy. *Ann Oto Rhinol Laryn*, 104(5):340–349, 1995.

S. Hertegård and J. Gauffin. Acoustic properties of the Rothenberg mask. *STL-QPSR*, 33(2-3):9–18, 1992.

S. Hertegård and J. Gauffin. Glottal area and vibratory patterns studied with simultaneous stroboscopy, flow glottography, and electroglottography. *J Speech Hear Res*, 38(1):85–100, 1995.

S. Hertegård, H. Larsson, and T. Wittenberg. High-speed imaging: applications and development. *Logoped Phoniatr Vocol*, 28:133–139, 2003.

M. Hirano. *Clinical Examination of Voice*. Springer-Verlag, Vienna, Austria, 1981.

M. Hirano, S. Kurita, and T. Nakashima. *Vocal fold physiology*, chapter The structure of the vocal folds, pages 33–41. University of Tokyo Press, Tokyo, Japan, 1981.

E. B. Holmberg, R. E. Hillman, and J. S. Perkell. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *J Acoust Soc Am*, 84(2):511–1787, 1988.

K. Ishizaka and J. L. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Sys Tech J*, 51(6):1233–1268, 1972.

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

S. Kiritani, H. Imagawa, and H. Hirose. Vocal cord vibration and voice source characteristics-observations by a high-speed digital image recording. *Proceedings of the 1st International Conference on Spoken Language Processing (ICSLP 1990)*, pages 61–64, 1990.

P. Kitzing. Clinical applications of electroglottography. *J Voice*, 4(3):238–249, 1990.

D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am*, 87(2):820–857, 1990.

A. K. Krishnamurthy. Glottal source estimation using a sum-of-exponentials model. *IEEE T Sig Proc*, 40(3):682–686, 1992.

A. K. Krishnamurthy and D. G. Childers. Two-channel speech analysis. *IEEE T Acoust Speech*, 34(4):730–743, 1986.

H. Larsson, S. Hertegard, P. Lindestad, and B. Hammarberg. Vocal fold vibrations: high-speed imaging, kymography, and acoustic analysis: a preliminary report. *Laryngoscope*, 110(12):2117–2122, 2000.

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

L. Lehto. *Occupational Voice–Studying Voice Production and Preventing Voice Problems with Special Emphasis on Call-Centre Employees*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 2007.

M. R. Mataušek and V. S. Batalov. A new approach to the determination of the glottal waveform. *IEEE T Acoust Speech*, 28(6):616–622, 1980.

P. Milenkovic. Glottal inverse filtering by joint estimation of an AR system with a linear input model. *IEEE T Acoust Speech*, 34(1):28–42, 1986.

R. L. Miller. Nature of the vocal cord wave. *J Acoust Soc Am*, 31(6):667–677, 1959.

R. B. Monsen and A. M. Engebretson. Study of variations in the male and female glottal wave. *J Acoust Soc Am*, 62(4):981–993, 1977.

P. J. Murphy. Spectral noise estimation in the evaluation of pathological voice. *Logoped Phoniatr Vocol*, 31(4):182–189, 2006.

I. R. Murray and J. L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J Acoust Soc Am*, 93(2):1097–1108, 1993.

A. Ní Chasaide and C. Gobl. Voice quality and f0 in prosody: Towards a holistic account. In *Proceedings of International Conference on Speech Prosody 2004*, pages 189–196, Nara, Japan, 2004.

M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE T Speech Audi P*, 7:569–586, 1999.

A. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *J Acoust Soc Am*, 49(1A):583–590, 1971.

T. D. Rossing. *The Science of Sound.* Addison-Wesley, 1990.

M. Rothenberg. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *J Acoust Soc Am*, 53(6):1632–1645, 1973.

K. R. Scherer. *The Neuropsychology of Emotion*, chapter Psychological Models of Emotion, pages 137–162. Oxford University Press, Oxford/New York, 2000.

K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Commun*, 40:227–256, 2003.

K. R. Scherer. Vocal affect expression: A review and a model for future research. *Psychol Bull*, 99(2):143–165, 1986.

H. Schlosberg. Three dimensions of emotion. *Psychol Rev*, 61:81–88, 1954.

M. Sondhi. Measurement of the glottal waveform. *J Acoust Soc Am*, 57:228–232, 1975.

K. Stevens. *Acoustic Phonetics*. MIT Press, 2000.

B. Story, I. Titze, and E. Hoffman. The relationship of vocal tract shape to three voice qualities. *J Acoust Soc Am*, 109(4):1651–1657, 2001.

B. H. Story. An overview of the physiology, physics and modeling of the sound source for vowels. *Acoustical Science and Technology*, 23(4):195–206, 2002.

B. H. Story and I. R. Titze. Voice simulation with a body-cover model of the vocal folds. *J Acoust Soc Am*, 97(2):1249–1260, 1995.

H. Strik, B. Cranen, and L. Boves. Fitting a LF-model to inverse filter signals. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech '93)*, volume 1, pages 103–106, Berlin, Germany, 1993.

H. W. Strube. Determination of the instant of glottal closure from the speech wave. *J Acoust Soc Am*, 56(5):1625–1629, 1974.

J. G. Švec and H. K. Schutte. Videokymography: High-speed line scanning of vocal fold vibration. *J Voice*, 10(2):201–205, 1996.

J. G. Švec, F. Šram, and H. K. Schutte. Videokymography in 2000: the present state and perspectives of the high-speed line-imaging technique. In *Proceedings of the 4th International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research*, pages 57–62, Jena, 2000.

R. Timcke, H. von Leden, and P. Moore. Laryngeal vibrations: measurements of the glottic wave. *Archiv Otolaryngol*, 68:1–19, 1958.

I. R. Titze. *Priciples of Voice Production*. Prentice-Hall, 1994.

I. R. Titze and J. Sundberg. Vocal intensity in speakers and singers. *J Acoust Soc Am*, 91(5):2936–2946, 1992.

J. W. van den Berg, J. T. Zantema, and J. P. Doornenbal. On the air resistance and the bernoulli effect of the human larynx. *J Acoust Soc Am*, 29(5):626–631, 1957.

D. E. Veeneman. and S. L. BeMent. Automatic glottal inverse filtering. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'84)*, 9:36.5.1–36.5.4, 1984.

R. Veldhuis. A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation. *J Acoust Soc Am*, 103(1):566–571, 1998.

E. Vilkman, E.-R. Lauri, P. Alku, E. Sala, and M. Sihvo. Effects of prolonged oral reading on F0, SPL, subglottal pressure and amplitude characteristics of glottal flow waveforms. *J Voice*, 13(2):303–315, 1999.

J. Walker and P. Murphy. A review of glottal waveform analysis. In Y. Stylianou, M. Faúndez-Zanuy, and A. Eposito, editors, *Progress in Nonlinear Speech Processing, Workshop on Nonlinear Speech Processing, WNSP 2005, Heraklion, Crete, Greece, September 20-23, 2005*, volume 4391 of *Lecture Notes in Computer Science*, pages 1–21. Springer Verlag, 2007.

D. Watson. *Mood and Temperament*. The Guilford Press, New York, NY, USA, 2000.

J. B. West. *Respiratory Physiology: The Essentials*. Lippincott Williams & Wilkins, seventh edition, 2000.

T. Wittenberg, M. Tigges, K. Spinnler, and U. Eysholdt. Some thoughts about 3D and stereo in laryngoscopy. In *Proceedings of the 4th International Workshop on*

Advances in Quantitative Laryngoscopy, Voice and Speech Research, pages 116–123, Jena, Germany, 2000.

D. Y. Wong, J. D. Markel, and A. H. Gray, Jr. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE T Acoust Speech*, 27(4):350–355, 1979.