Tools for analysing the voice

Developments in glottal source and voice quality analysis

Thesis submitted for the degree of Doctor of Philosophy

By

John Kane
Supervisor: Prof. Christer Gobl

Phonetics and Speech Laboratory
School of Linguistic, Speech and Communication Sciences
Trinity College Dublin

Declaration

I, the author, declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signature of author: _____

John Kane
28th September 2012

For my two special people.

Pain is temporary, failure lasts forever.

—Lance Armstrong

Summary

This thesis documents a range of research carried out on the topic of glottal source and voice quality analysis. Initially, a review is given of the physiological and acoustic correlates of different vocal settings. This is followed by a discussion of the importance of glottal source and voice quality variation in spoken communication, and the impact of modelling these aspects on speech technology. Despite the potential benefit of acoustic characterisation of the glottal source for speech technology, existing algorithms often suffer from a lack of robustness. To address this, the present thesis describes and evaluates a set of novel algorithms aimed at improving the robustness. The algorithms come under two headings: fine-grained, glottal synchronous methods and coarse-grained, voice quality detection methods. In terms of fine-grained methods a new algorithm, SE-VQ, has been developed which is optimised for analysis of a range of voice qualities. While maintaining the precision of the state-of-the-art on neutral speech, the new algorithm is shown to significantly improve performance on creaky voice regions. SE-VQ is then utilised as part of a novel LF model based parameterisation method (DyProg-LF) of estimated glottal source signals. The dynamic programming algorithm used in DyProg-LF is shown to avoid the common problem of inconsistencies in parameter trajectories and is shown to provide better parameterisation than the state-of-the-art on both a carefully controlled dataset with manually obtained reference values as well as on a larger speech dataset. For coarse-grained methods, a new parameter, the Maxima Dispersion Quotient (MDQ), is proposed for discriminating breathy to tense voice. MDQ was shown to outperform existing parameters for discriminating the voice qualities, particularly for continuous speech, and also in terms of robustness to additive noise. A new method for detecting creaky voice is also described which utilises two parameters derived from the Linear Prediction-residual signal. These parameters are used as input features to a decision tree classifier which is shown to significantly outperform the state-of-the-art on a range of speech data varying in terms of speaker, gender, language, recording condition and speaking style. Finally, a software package, the *Voice analysis toolkit*, which contains the algorithms developed as part of this thesis, has been made publicly available. This has been done to encourage usage of the newly developed algorithms in applied work and future algorithm evaluations.

Acknowledgments

Quite a number of people, in fact more than can be mentioned here, have helped me through the last few years and I would like to take this opportunity to offer my thanks to them.

First off, I would like to sincerely thank my supervisor, Christer Gobl, who sparked my interest in speech processing during my masters and who has truly been the ideal supervisor throughout the subsequent four years of my Ph. D. Rarely can a supervisor strike that fine balance between supervision and autonomy which allows one to explore and experiment while still ensuring the core research path is kept. I would also like to thank his wife, Ailbhe Ní Chasaide, who has been at times like a second supervisor to me and has made me feel completely at home in lab.

My gratitude to Science Foundation Ireland, who funded me through both the CNGL and FastNet projects despite the present economic difficulties in Ireland.

I would like to thank everyone at the Phonetics and Speech Lab, especially for tolerating my distractions and anecdotes, and more importantly for generously offering both academic and personal support. This thanks also extends to everyone at the Speech Communication Lab and at the UCD Muster group. Having a good group of us suffering similar plights has certainly helped maintain some degree of sanity.

A big thank you to all the people I have collaborated with over the last four years. In particular: Stefan, Tom, Tuomo, Helena, Kinga, João, Mark, Irena, Catha and Nick, I have learnt so much from you that I would never have been able to learn by myself in this short space of time.

To my friends, my mother and brother your support has always been there throughout the Ph. D., and besides your love and encouragement, you have been there to remind me that there is more to this world than research. Thank you so much.

Most importantly I would like to thank my two special people, Áine and Luadh. Luadh why you decided to start sleeping all through the night for the last leg of the Ph. D. I have no idea, but your brilliant, true smile was at times the only thing that could brighten up an otherwise bleak and miserable day. You will never realise the joy you have brought me. And to Áine, my love, that while doing your own Ph. D. in parallel you could still find the energy to cheer me up and get me out of a rut has meant I have made it this far with my body and my mind still in tact. For explaining what you have meant to me, no amount of gesture could ever suffice.

John Kane

September, 2012

Contents

Notation

f_0	:	fundamental frequency
fs	:	sampling frequency of a digital signal
$s(n)$	:	digital speech signal
$S(\omega)$	:	Speech spectrum of $s(n)$
$S(z)$	:	Z-transform of $s(n)$
$V(z)$	:	Z-transform of the vocal tract filter
$R(z)$	:	Z-transform of the effect of lip radiation
$G(z)$	:	Z-transform of the glottal source
$\hat{s}(n)$	:	Complex cepstrum of speech signal $s(n)$
$g(n)$	:	glottal source
$g'(n)$	:	glottal source derivative
$g_{LF}(n)$	:	Synthetic LF model signal of the glottal source
$r_{LP}(n)$	:	Linear Prediction residual
w	:	window function

List of Abbreviations

ARX	:	Auto-Regressive eXogenous
CALM	:	Causal Anti-causal Linear Model
CCEPS	:	Complex Cepstrum-based decomposition
CPIF	:	Closed-Phase Inverse Filtering
EGG	:	Electroglottographic signal
DAP	:	Discrete all-pole model
DYPSA	:	DYnamic programming projected Phase Slope Algorithm
DEGG	:	Derivative electroglottographic signal
FAR	:	False Alarm Rate
FFT	:	Fast Fourier Transform
GCI	:	Glottal Closure Instant
GOI	:	Glottal Opening Instant
GUI	:	Graphical User Interface
HMM	:	Hidden Markov Model
HSD	:	Honestly Significant Difference test
HRF	:	Harmonic Richness Factor
HTS	:	HMM-based speech synthesis system
IAIF	:	Iterative Adaptive Inverse filtering
IDA	:	Identification Accuracy
IFP	:	Intra-Frame Periodicity
IPS	:	Inter-Pulse Similarity
IR	:	Identification Rate
LF	:	Liljencrants-Fant
LoMA	:	Lines of Maximum Amplitude
LTI	:	Linear Time-Invariant
LP	:	Linear Prediction
LPC	:	Linear Predictive Coding

MDQ	:	Maxima Dispersion Quotient
NAQ	:	Normalised Amplitude Quotient
OQ	:	Open Quotient
MFCCs	:	Mel-Frequency Cepstral Coefficients
MSPD	:	Mean Squared Phase Difference
MR	:	Miss Rate
SEDREAMS	:	Speech Event Detection using the Residual Excitation and A Mean-based Signal
SFI	:	Science Foundation Ireland
SNR	:	Signal-to-Noise Ratio
SWT	:	Stationary Wavelet Transform
TTS	:	Text-To-Speech synthesis
YAGA	:	Yet Another GCI/GOI Algorithm
ZFF	:	Zero Frequency Filtering
ZZT	:	Zeros of the Z-Transform

List of Figures

List of Tables

Chapter 1

Introduction

The human voice is perhaps the most powerful and ubiquitous mechanism for communicating that exists. It is used for a wide variety of functions in spoken interaction from the signalling of prominence in an utterance to the expression of affective states and attitudes. The research documented in this thesis focuses on the development of analysis tools for the modelling and characterisation of targetted aspects of the voice. Specifically, this thesis is concerned with the effect of the glottal source, emanating from vibration of the vocal folds, and its impact on the speech signal. Furthermore, this research is concerned with the variation and dynamics of the glottal source contribution and its effect on perceived voice quality. Note that the research does not extend to singing or disordered voices but instead is dedicated to the effective characterisation of the glottal source and voice quality variation in non-pathological speech.

Accurate acoustic characterisation and modelling of the glottal source is desirable for a range of applications including linguistic analysis tools, and as acoustic features used in speech technology (e.g., speech synthesis, voice modification, speaker identification). However, the potential of acoustic features relating to the glottal source and voice quality has yet to be fully exploited in these areas. This is largely due to a perceived lack of robustness in automatic analysis methods.

It follows that the purpose of this thesis is to develop more robust algorithms for analysing different aspects of the glottal source contribution. The analysis methods developed and evaluated come under two main headings: *fine-grained methods* and *coarse-grained methods*. *Fine-grained methods* refers to the analysis at the level of individual glottal pulses produced by vocal fold vibration at the level of the glottis. A new algorithm is proposed for detecting glottal closure instants (GCIs) suitable for a wide range of voice types. This is then used as part of a novel method for automatic parameterisation of the

glottal source, which is designed to simulate the strategies used in labour-intensive manual analysis. There then follows a comprehensive empirical evaluation of glottal source-vocal tract filter decomposition as well as glottal source parameterisation methods. *Coarse-grained methods*, on the other hand, refers to the detection of perceivable changes in voice quality given rise to by substantial changes in phonation type. Novel algorithms are proposed for differentiating breathy-to-tense voice, as well as for detecting creaky voice. Furthermore, these algorithms are evaluated in terms of robustness to degraded conditions.

The algorithms described in these two parts of the thesis, although distinct from one another, can be used to provide complementary descriptions and modelling of the voice. For the *fine-grained methods* one can obtain precise detail on the changes in the glottal source from pulse-to-pulse. However, these methods typically require high quality recording conditions. At the same time it is widely believed that speakers vary their vocal timbre considerably more in natural conversational settings, which are often likely to be recorded in less than ideal conditions. The approaches described in the *coarse-grained methods* part of the thesis are, hence, designed for the purpose of detecting changes in voice quality and evaluated in terms of robustness to simulations of degraded recording conditions. As a result these approaches can be used to study and model naturally occurring changes in the voice.

A concrete example can be useful to illustrate how the combination of these two approaches could be exploited in terms of a specific application. One can consider the speech technology application of statistical parametric synthesis. New synthesis platforms have been developed which involve the use of a glottal source model which can be altered in order to approximate an alternative voice quality. However, two problems exist. On the one hand, there are known problems with the automatic glottal source modelling used in these applications. The algorithms described in the *fine-grained methods* part may contribute to improving glottal source modelling, and potentially providing a more natural rendering of the speaker's vocal timbre. On the other hand, although these new platforms have been developed which facilitate flexibility in terms of the voice, it is not clear how exactly the voice should be varied, and for what communicative function. A simple linear alteration of the glottal source model shape is unlikely provide much value. By exploiting the algorithms described in the *coarse-grained methods* part, one could study and model how speakers dynamically vary their voice quality in natural conversation settings. This information could then be used to instruct how the glottal source should be varied for particular communicative functions. Such an approach may be useful for developing expressive and conversational parametric speech synthesis.

The methods described in this thesis are, however, not designed solely for any one spe-

cific application and it is envisaged that they may be suitable for a broad range of uses. Nevertheless, the development of these approaches has not been completely without consideration of potential uses. To illustrate this, each experimental chapter which proposes a novel analysis method (i.e. Chapters 4 - 5 and 7 - 8) concludes with a short description of how these methods could be applied and, in many cases, how these methods have already been exploited in speech applications following collaboration with fellow researchers. Furthermore, in order to encourage use of the developed algorithms, the Matlab code for the various methods has been made available online[1] in the form of a *Voice analysis toolkit*.

The thesis is organised as follows: The next two chapters provide a comprehensive review of the literature of this area, in terms of the physiology and acoustics of speech and then of the importance of the glottal source in spoken communication and speech technology. These two chapters make up the first part of the thesis. The second part, *fine-grained methods*, begins with a study on the development of glottal closure instant (GCI) detection for a range of voice qualities (Chapter 4). There then follows a chapter describing a new method for automatically parameterising the glottal source (Chapter 5). This thesis part is concluded with a large evaluation of the different automatic methods used in glottal source analysis (Chapter 6). The third part, *coarse-grained methods*, has an initial chapter describing and evaluating a method for differentiating breathy-to-tense voice qualities by utilising features of the wavelet transform (Chapter 7). The next chapter describes an approach for detecting creaky voice in speech signals (Chapter 8). The final part of the thesis provides a general discussion of the findings in this thesis (Chapter 9). The potential impact of the developed methods is described and the future directions of this line of research are outlined.

Contributions of this thesis

A summary of the main contributions of this thesis is now given.

1. **Glottal closure instant detection algorithm suitable for a range of voice qualities:** In Chapter 4 a new algorithm is proposed for automatically detecting glottal closure instants (GCIs) on a range of voice qualities. Most studies on GCIs tend not to focus on evaluation on speech data containing high variability of voice quality. However, the glottal closing characteristics of different voice qualities can significantly affect the performance of GCI algorithms. The proposed method applies a dynamic programming algorithm to help improve the selection of Linear Prediction

[1] `https://github.com/jckane/Voice_Analysis_Toolkit`

(LP) residual peaks. This is particularly important when there are no prominent LP-residual peaks as is often the case in breathy and harsh voice. Furthermore, a post-processing procedure is used to remove false positives which often occur in creaky voice.

2. **LF model fitting algorithm based on dynamic programming:** In Chapter 5 an algorithm is described for fitting LF model pulses to an estimated glottal source signal. One common problem for model fitting algorithms is the consistent setting of the glottal opening instant. The proposed algorithm exploits a dynamic programming algorithm in order to avoid sudden changes in the model settings in speech regions with relatively high stationarity. The setting of the parameters of the dynamic programming algorithm was done following analysis of the strategies used on reliable reference data obtained by manual analysis. The new algorithm was shown to perform favourably compared to two comparison method both in terms of a quantitative evaluation and more qualitative one.

3. **Novel measurement for differentiating breathy-to-tense voice:** A parameter for differentiating breathy to tense voice is proposed in Chapter 7 based on features of a wavelet transform. Maxima derived following wavelet decomposition are often used for detecting edges in image processing, where locations of these maxima organise in the vicinity of the edge location. Similarly for tense voice, which typically displays sharp glottal closing characteristics, maxima following wavelet analysis are organised in the vicinity of the glottal closure instant (GCI). Contrastingly, as the phonation type tends away from tense voice towards a breathier phonation it is observed that the maxima become increasingly dispersed. The proposed parameter is designed to measure the extent of this dispersion and is shown to compare favourably to existing parameters, particularly for analysis of continuous speech.

4. **Novel algorithm for automatically detecting creaky voice:** In Chapter 8 a method for automatically detecting regions of creaky voice in a speech signal is described and evaluated. The detection of creaky voice is obtained using two parameters, which are derived from the LP-residual input through a resonator, as input feature to a binary decision tree classifier. The method is evaluated on a range of speech data varying in terms of speaker, gender, language, recording condition and speaking style, and is shown to significantly outperform the state-of-the-art.

5. **Voice analysis toolkit:** A final contribution of this thesis is a toolkit containing the above algorithms which has been made publicly available. This has been done to

encourage usage of the methods described in the thesis and to exploit any subsequent feedback to refine and improve these methods. The README file for the toolkit is given in Appendix B.

1.1 Authors publications

Journal publications

- Kane, J., Gobl, C., (2013) Evaluation of glottal closure instant detection in a range of voice qualities, *Speech Communication*, 55(2), pp. 295-314.
- Scherer, S., Kane, J., Gobl, C., Schwenker, F., (2013) Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification, *Computer Speech and Language* 27(1), pp. 263-287.
- Kane, J., Drugman, T., Gobl, C., (2013) Improved automatic detection of creak, *Computer Speech and Language* 27(4), pp. 1028-1047.
- Kane, J., Gobl, C., (2013) Automating manual user strategies for precise voice source analysis, *Speech Communication* 55(3), pp.397-414.
- Kane, J., Gobl, C. (2013) Wavelet maxima dispersion for breathy to tense voice discrimination, *IEEE Transactions on Audio, Speech and Language processing* 21(6), pp. 1170-1179.

Selected conference publications

- Scherer, S., Kane, J., Gobl, C., Schwenker, F., (2013) The effect of fuzzy training targets on voice quality classification, *Workshop on: Multimodal pattern recognition of social signals in human computer interaction, Tsukuba, Japan.*
- Drugman, T., Kane, J., Gobl, C., (2012) Resonator based creaky voice detection, *Proceedings of Interspeech, Portland, Oregon, USA.*
- Drugman, T., Kane, J., Gobl, C., (2012) Modeling the creaky excitation for parametric speech synthesis, *Proceedings of Interspeech, Portland, Oregon, USA.*
- Kane, J., Yanushevskaya, I., Ní Chasaide, A., Gobl, C., (2012) Exploiting time and frequency domain measures for precise voice source parameterisation, *Proceedings of Speech Prosody, Shanghai, China.*, 143-146.

- Székely, É, Kane, J., Scherer, S., Gobl, C., Carson-Berndsen, J., (2012) Detecting a targetted voice style in an audiobook using voice quality features, *Proceedings of ICASSP, Kyoto, Japan,* 4593-4596.

- Scherer, S., Layher, G., Kane, J., Neumann, H., Campbell, N., (2012) An audiovisual political speech analysis incorporating eye-tracking and perception data, *Proceedings of LREC, Istanbul, Turkey,* 1114-1120.

- Kane, J., Gobl, C. (2011) Identifying regions of non-modal phonation using features of the wavelet transform, In *Proceedings of Interspeech 2011, Florence, Italy*, 177-180.

- Cabral, J., Kane, J., Gobl, C., Carson-Berndsen, J., (2011) Evaluation of glottal epoch detection algorithms on different voice types, *Proceedings of Interspeech, Florence, Italy,* 1989-1992.

- Kane, J., Pápay, K., Hunyadi, L., Gobl, C., (2011) On the use of creak in Hungarian spontaneous speech, *Proceedings of ICPhS, Hong Kong.*

- Kane, J., Kane, M., Gobl, C., (2010) A spectral LF model based approach to voice source parameterisation, *Proceedings of Interspeech, Makuhari, Japan,* 2606-2609.

- Yanushevskaya, I., Gobl, C., Kane, J., Ní Chasaide, A., (2010) An exploration of voice source correlates of focus, *Proceedings of Interspeech, Makuhari, Japan,* 462-465.

- Kane, J., Gobl, C. (2009). Automatic parameterisation of the glottal waveform combining time and frequency domain measures, in *Proceedings of 6th Maveba International Workshop, Florence.*

Part I

Literature review

Chapter 2

Production, acoustics and acoustic modelling of speech

2.1 Speech production

The production of speech is often described in terms of three physiological components: the respiratory system, the larynx and the supralaryngeal vocal tract (Lieberman and Blumstein, 1988, Chap. 2). The remainder of this section will provide a brief overview of physiological and acoustic aspects of these separate components. As this thesis is primarily concerned with the phonatory aspects of speech, a longer amount of time will be spent discussing the larynx including a review of the physiological mechanisms of a range of phonation types.

2.1.1 The respiratory system

All aspects of speech require the production of an airstream at the lungs which is then modified by different parts of the vocal system (Hardcastle, 1976). Inspiration typically involves a sudden increase in lung volume, followed by a relatively constant decrease during expiration. During the rapid inspiration phase the external intercostal muscles and the diaphragm aid the expansion of the lungs (Titze, 1994). Titze (1994) outlines three phases during expiration. In the first stage the outward movement of the lungs and the ribs (so called elastic-recoil) controls the lung pressure. Next the intercostal muscles are used to put pressure on the lungs as the strength of the elastic-recoil reduces. The final phase also involves the use of the intercostal muscles and often in combination with the back muscles to force further air out (Hoit and Hixon, 1986).

This repetitive process provides the airflow which is subsequently modified by the rest of the vocal apparatus during speech.

2.1.2 The larynx

The first structure to modify the flow of air passing out of the lungs is the larynx (Hardcastle, 1976). The larynx is located in the neck and consists of soft tissue protected by cartilage. The different laryngeal cartilages and muscles are shown in Figure 2.1 with a schematic illustration of the muscular tensions relevant to speech in Figure 2.2. The vocal folds are located within the larynx and the orifice between them is called the *glottis*. The opening and closing of the glottis is largely determined by the settings of the laryngeal musculature. A summary of the main muscular tensions exploited by speakers during phonation is given with reference to Figures 2.1 and 2.2:

- *Adductive tension*: is caused by the tension between the oblique and transverse interarytenoid muscles (see the top right panel in Figure 2.1) which pulls the arytenoids together (Gobl, 1989).
- *Medial compression*: controls the closure of ligamental glottis (see bottom right panel of Figure 2.1). Medial compression is controlled by the lateral cricoarytenoid muscle and also by the external thyroarytenoid muscle. In order to achieve closure of the cartilaginous glottis however, there also needs to be *adductive tension* (Laver, 1980).
- *Longitudinal tension*: is the result of contraction of the vocalis and the cricothyroid muscles which controls the tension of the vocal folds (Laver, 1980).

The myoelastic-aerodynamic theory, described by van den Berg (1958), has often been used to explain the vibration of the vocal folds. Starting from a shut position, with the vocal folds held together by adductive muscular tension, they are then forced apart due to an increase in subglottal pressure (Hardcastle, 1976). As the air passing through the glottis increases in speed, a pressure difference is created, known as the Bernoulli effect. This is combined with the elastic forces of the vocal fold tissue which 'sucks' the vocal folds together. However, it has been pointed out that the myoelastic-aerodynamic theory is not sufficient to explain how the vocal fold oscillation is self-sustained and further explanation on this are given in Titze (1994).

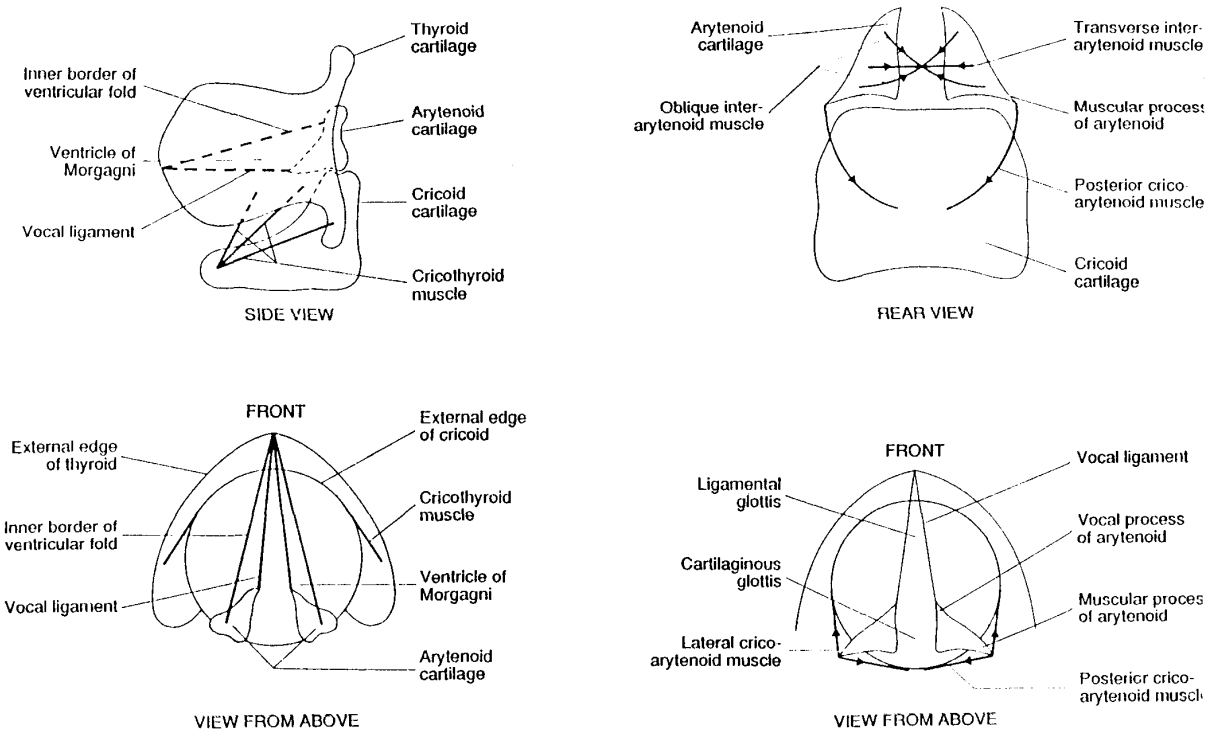


Figure 2.1: Illustration of the larynx from the side, the rear and above with relevant laryngeal muscles labelled (based on illustration in Laver, 1980)

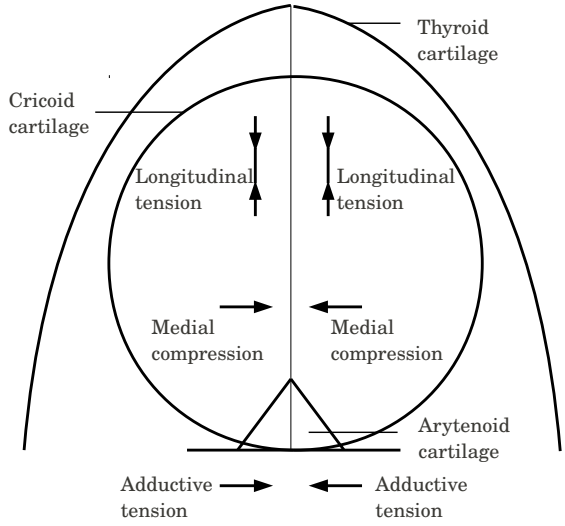


Figure 2.2: Illustration of the three main dimensions of laryngeal muscular tension (based on illustration in Laver, 1980)

Voice quality and phonation types

The use of terminology relating to voice quality varies considerably in the literature. To minimise potential confusion, the terms and definitions used in this thesis are, as far as possible, consistent with the descriptions in Laver (1980). Thus, voice quality refers to the auditory colouring of a persons voice, although this research is primarily concerned with voice quality variation brought about by changes in laryngeal activity. Changes in the tension settings of the laryngeal musculature will give rise to different *phonation types*. The term *modal voice* is used to depict a neutral phonation type, involving moderate levels of laryngeal tension, periodic vocal fold vibration with a minimum of pulse-to-pulse irregularities, efficient glottal excitation with full or essentially full glottal closure, and no audible frication noise (Gobl and Ní Chasaide, 1992). A phonation type deviating considerably from this description will be considered *non-modal*.

Laver (1979) describes a speaker's voice quality as being composed of the longer-term setting of the vocal system combined with dynamic shifts in the system used for communicative purposes. Ladefoged (1960) initially used the term *organic* when referring to the individuals unique physical features that affects the quality of their voice, e.g., vocal tract length, size of tongue, shape of laryngeal structures, etc. Voice quality, as a result of these organic features, is then varied as part of the person's speaking habits (Abercrombie, 1967). An overview of the laryngeal physiology involved in different phonation types is now given along with a description of the resulting acoustic characteristics. A description of the use of voice quality in speech communication given in Chapter 3.

Modal voice

Modal voice is the type of phonation which is effectively normal [1] sounding phonation. Catford (1964) describes modal voice as having both the ligamental and cartilaginous glottis "functioning as a single unit".

It is described as being the most 'efficient' form of phonation produced using moderate adductive tension, medial compression and longitudinal tension. Increased longitudinal tension is mainly used to increase pitch (van den Berg, 1968). The vibration of the vocal folds is often quasi-periodic, with minimal frication and full glottal closure (Gobl, 1989). As stated previously, other phonation types will be described in reference to this description of modal voice. The sequence of phonation types will roughly follow a lax to tense continuum starting with whisper/whispery voice. Falsetto, which does not fit clearly

[1] The use of 'normal' is discouraged as it suggests that other voice qualities are abnormal (Laver, 1980, p. 109)

into this sequence will be described last.

Whisper and whispery voice

Whisper is typically reported to involve a triangular opening of the glottis demonstrating an upside-down Y shape (Pressman, 1942; Luchsinger and Arnold, 1965; Laver, 1980). Achieving this shape is thought to involve low adductive tension and moderate to high levels of medial compression (Laver, 1980). The constant flow of air through this constriction and the resulting turbulence produces the perceived quality of whisper (van den Berg, 1968). Studies in the literature often discriminate types of whisper based on vocal effort (Monoson and Zemlin, 1989; Solomon et al., 1989); e.g., *True* whisper (low-effort) and *stage* whisper (high-effort) usually produced by actors when wishing to portray the form of whisper while being sufficiently loud to be heard by an audience (Obin, 2012).

Whisper can be combined with modal voice to create the compound *whispery voice* (Laver, 1980). In whispery voice although the triangular opening is thought to be maintained, the ligamental portion of the glottis closes during each glottal cycle. Note that in whisper there is a complete absence of periodic vocal fold vibration, whereas in whispery voice there is indeed periodic vibration along a portion of the vocal fold length.

Figure 2.3 shows a broadband spectrogram of the same utterance produced by a male speaker in modal voice (top panel) and in whisper (bottom panel). The lower formant patterns of the two utterances are clearly similar. However, for the modal utterance one can observe vertical lines corresponding to the individual glottal pulses which are not present for the utterance produced with whisper.

Breathy voice

Breathy voice is perhaps the most studied non-modal phonation type (Hanson et al., 2001). During breathy phonation the vocal folds vibrate in a more inefficient [2] manner and the vibration is accompanied by audible frication. In terms of the muscular tension there is believed to be minimal adductive tension and low medial compression (Laver, 1980). The glottis can undergo incomplete closure as a result of reduced muscular tension, which can allow a constant jet of turbulent air to pass through the glottis. This incomplete closure usually manifests itself as a gap in the posterior portion of the vocal folds (Zemlin, 1964). Although the size of this gap reduces as the phonation type moves in the direction of tense voice, there may still be a gap during speech where there is no audible sensation of

[2] Inefficient here refers to the need for increased effort required to produce the same amount of power in their speech signal

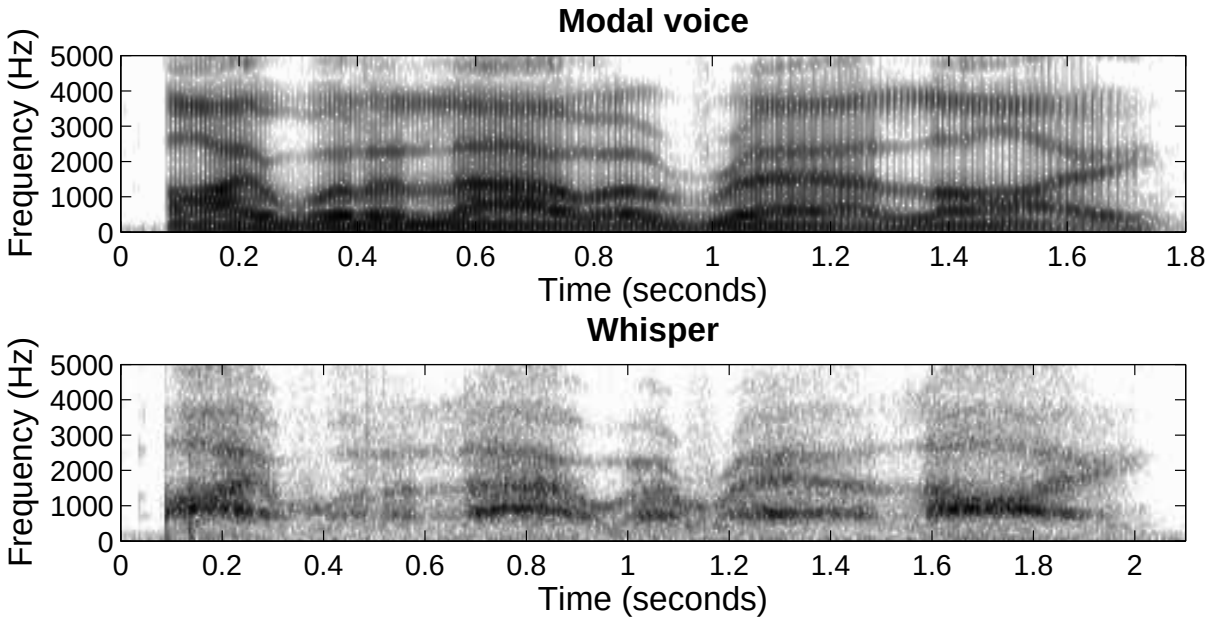


Figure 2.3: Broadband spectrogram of the utterance *I will allow a rare lie* produced in modal voice (top panel) and whisper (bottom panel) by a male speaker.

aspiration noise (Chen et al., 2011).

When breathiness is perceptually noticeable the speech signal typically has certain acoustic attributes and Klatt and Klatt (1990) outlined some of these main features. One feature is increased first formant (F1) bandwidth. This is likely to be due to increased losses at the glottis during the open phase of the glottal cycle (Fant, 1979). A rough correlate of F1 bandwidth, suggested in Hanson (1997), involves subtracting the amplitude of the harmonic closest to F1 from the speech spectrum from the amplitude of the first harmonic from the differentiated glottal source spectrum (i.e. H1*-A1).

Another feature of breathiness, as a result of non-simultaneous closure along the length of the glottis (Laver, 1980), is the impact on the spectral tilt in the mid to high frequencies (Hanson, 1997). Typically breathiness will produce steeper spectral tilt, compared with modal voice. Hanson (1997) suggest an acoustic correlate of this by subtracting the amplitude of the harmonic closest to the third formant, taken from the differentiated glottal source spectrum, from the amplitude of the first harmonic, also from the differentiated glottal source spectrum (i.e. H1*-A3*).

Further evidence of breathiness in a speech signal comes from the turbulent air produced at the glottis. Klatt and Klatt (1990) stated that if a speech signal is bandpass filtered in the third formant region this noise feature becomes visually apparent in the speech waveform and suggested a subjective four point rating scale for describing this.

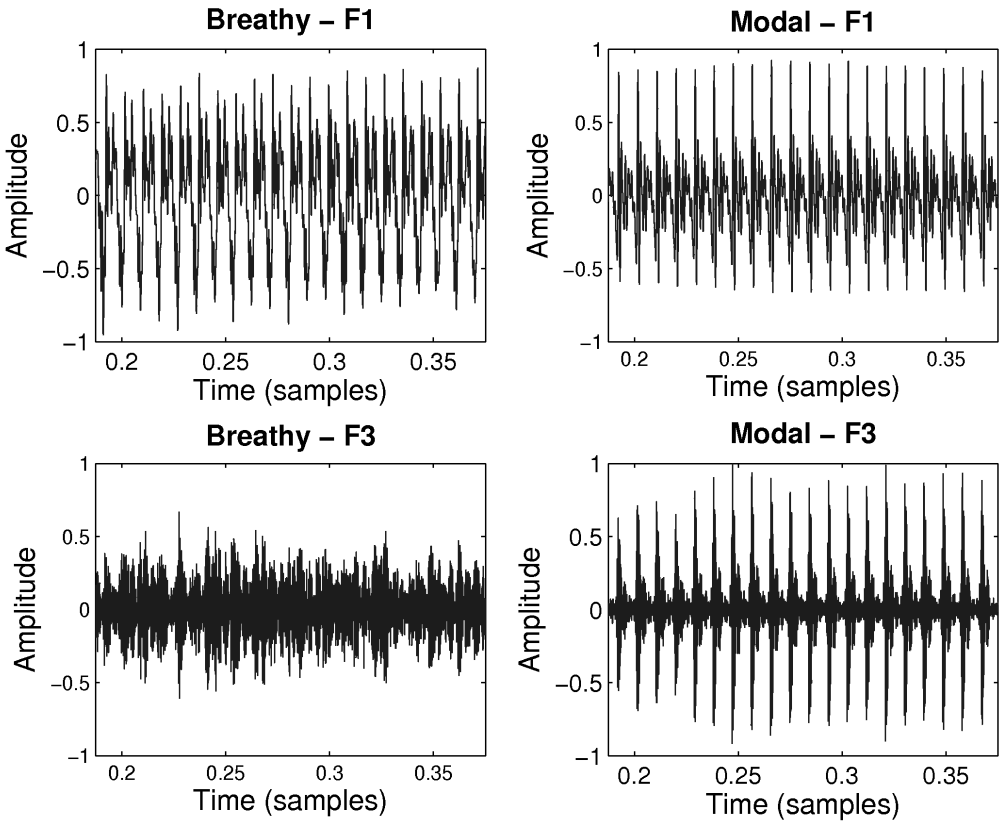


Figure 2.4: Breathy (left column) and modal (right column) /a/ vowel produced by a male speaker and bandpass filtered in the first formant (top row) and third formant (bottom row) regions.

This observation is illustrated in Figure 2.4, where both breathy and modal vowels bandpass filtered in the first formant region show periodic behaviour. Contrastingly, for the breathy vowel bandpass filtered in the third formant region the outputted waveform is extremely noisy, compared to the highly periodic filtered modal vowel. Researchers have since attempted to formalise this measurement. One notable attempt is the use of a three point shimmer measurement (Amplitude Perturbation quotient) on the bandpass filtered signals (Ito, 2004). A more recent approach to this has involved comparing synchronicity of the amplitude envelopes of the speech signal bandpass filtered at around F1 and one bandpass filtered around F3 through the use of cross-correlation (Ishi et al., 2010).

The perceptual importance of acoustic correlates of breathy voice were investigated Gobl and Ní Chasaide (1999b) where it was found that spectral tilt was the main factor contributing to the perception of breathiness. A perhaps unexpected finding in that study, particularly given observations in the previous paragraph, was that the aspiration noise parameter in the KLSYN88 synthesiser (Klatt, 1980) contributed little to the perception

of breathiness. However, given the difficulty of modelling aspiration noise, particularly in terms of the frequency and time domain modulation of a noise source, this finding may be more due to the model of aspiration noise in the KLSYN88 synthesiser and not aspiration noise in human speech production.

Tense voice

The above voice qualities were described in terms of their laryngeal physiology as phonation was the main factor in producing the perceived voice quality. Tense voice, however, is generally thought of as involving elevated tension settings throughout the entire vocal system (Laver, 1980). At the laryngeal level, the increased tension may sometimes result in a phonation type which can be adequately described as harsh voice. However, compared to harsh voice, the increase in laryngeal tension in tense voice is typically less extreme, and may not give rise to the characteristic irregularities in vocal fold vibration associated with harsh voice (Laver, 1980). Hence, tense voice is here used to refer to a voice quality produced by an increase in the tension settings compared to modal voice, but which does not display the irregular vocal fold vibration associated with harsh voice.

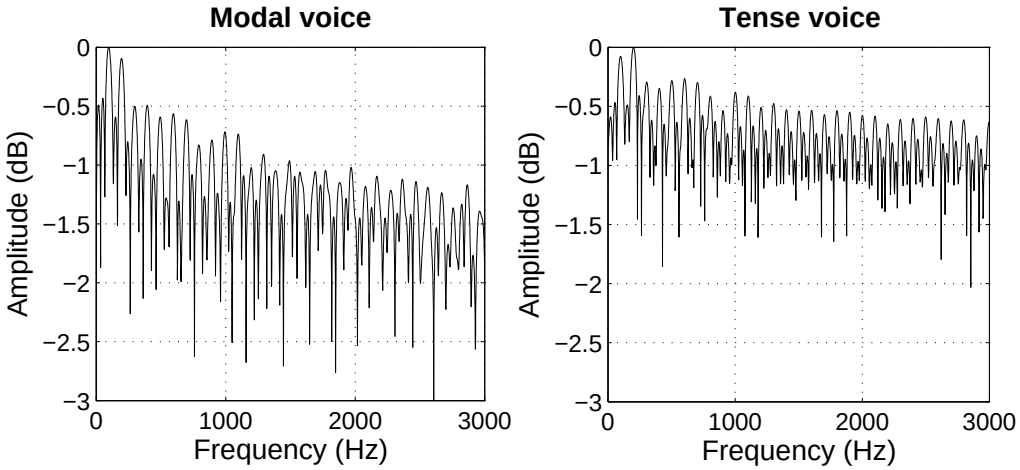


Figure 2.5: Amplitude spectrum of the differentiated glottal source signal, estimated by inverse filtering, of the same vowel produced in modal voice (left panel) and tense voice (right panel) by a male speaker. Note that the amplitude of the two spectra has been normalised to facilitate visual comparison of the spectral roll-off.

In terms of the acoustic characteristics, tense voice typically displays a less steep spectral slope (this can be observed in differentiated glottal source spectra for modal and tense voice in Figure 2.5). The spectrum displays a stronger harmonic richness with more partials exceeding the noise level in the higher frequencies than in modal voice (Gobl and Ní

Chasaide, 1992). Also, as can be seen in Figure 2.5, the amplitude of the second harmonic can be higher than that of the first (Hanson, 1997).

Harsh voice

Harsh voice is thought to be a result of excessive vocal fold tension and is often produced with a low pitch (Van Riper and Irwin, 1958; Zemlin, 1964). There is likely to be extreme adductive tension and medial compression (Brackett, 1940), essentially over contraction of the muscle tensions involved in modal voice (Laver, 1980). Harsh voice, however, is believed to involve other aspects of the larynx than just the glottis. Esling and Harris (2003), following inspection of laryngoscopic images, suggest that, in the 21 tone of Bai, constriction of the laryngeal sphincter in the epilaryngeal tube can be applied, sometimes with aryepiglottic trilling. Note that above the ventricular folds (see side view in Figure 2.1) is where the aryepiglottic folds are located. Harsh voice is sometimes referred to as ventricular voice, when produced with a high pitch.

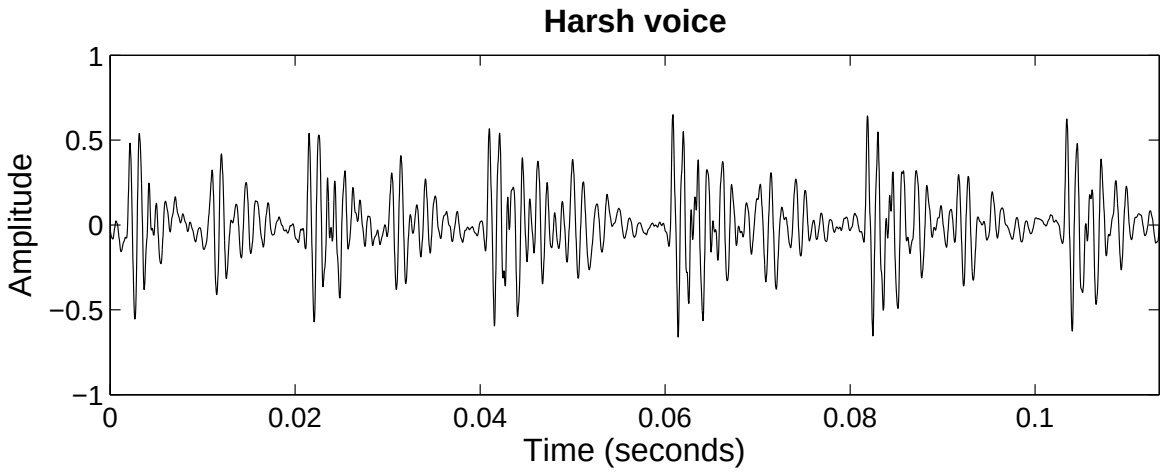


Figure 2.6: Speech waveform of a segment of an utterance by a male speaker produced with harsh voice.

Harsh voice is known to display irregularities in terms of amplitude and frequency of successive glottal pulses (Wendahl, 1964; Laver, 1980; Ishi et al., 2008a), often referred to as diplophonia. Such irregularities in amplitude are illustrated in Figure 2.6. In particular these pulse to pulse fluctuations in amplitude are thought to affect the 'rough' sensation of harsh voice (Wendahl, 1964). This modulation in amplitude may contribute to distorting the perception of pitch in harsh voice (Warren, 1982). The strong presence of noise in the speech spectrum is also thought to be characteristic of harsh voice (Fairbanks, 1960).

Creaky voice

Creaky voice is sometimes referred to as *vocal fry*, *glottal fry* or *laryngealisation*. It has been suggested that creaky phonation involves strong adductive tension and medial compression and low longitudinal tension (Fønagy, 1962) as well as low subglottal pressure (Monsen and Engebretson, 1977), compared to modal voice. It is typically produced with a lower fundamental frequency. Ladefoged (1971) states that during creaky phonation only the anterior parts of the vocal folds vibrate while the posterior parts are held together.

Some further insights into the physiology involved in creaky voice production were highlighted in Edmondson and Esling (2006), including the occurrence of *ventricular incursion*. Ventricular incursion is when the ventricular folds push down and cover the *true* vocal folds, causing an increased mass and, as a result, lowers the frequency of vibration (Moisik and Esling, 2011). This ventricular incursion can also result in secondary vibrations occurring above the glottis.

An often referred to, impressionistic description of the perception of creaky voice is given by Catford (1964): " a rapid series of taps, like a stick being run along a railing". However, the auditory criterion for creaky voice applied in this thesis is: "a rough quality with the sensation of additional impulses", which comes from Ishi et al. (2008b).

Many of the resulting acoustic characteristics of creaky voice are clearly distinct from modal voice. One of these features is the very long glottal pulse duration (where pulses can occasionally be as long as 100 ms, see Hollien and Wendahl, 1968; Blomgren et al., 1998). Such findings are corroborated by the results of psychoacoustic experiments carried out in Titze (1994), which demonstrated that human listeners begin to perceive individual pulses from around 70 Hz. Another acoustic feature often reported is the presence of secondary excitations, shown in electroglottographic (EGG) and speech pressure signals (Blomgren et al., 1998) as well as in glottal source signals estimated by inverse filtering (Gobl and Ní Chasaide, 1992). These secondary excitations may perhaps be explained by the occurrence of ventricular incursion, mentioned above. A further observation is that there is little or no superposition of formant oscillations between adjacent glottal pulses (Ishi et al., 2008b). One can frequently observe that oscillations from the vocal tract resonances have almost completely decayed before the start of the next pulse.

Note that different temporal excitation patterns can result in the perception of creaky voice. In certain cases, the fundamental frequency drops below a certain auditory threshold and reasonably periodic vibration is maintained. In other cases, the periodicity is highly irregular, often involving spells of diplophonia. In order for the perception of creaky voice, as opposed to harsh voice, a certain amount of glottal pulses need to have a very low

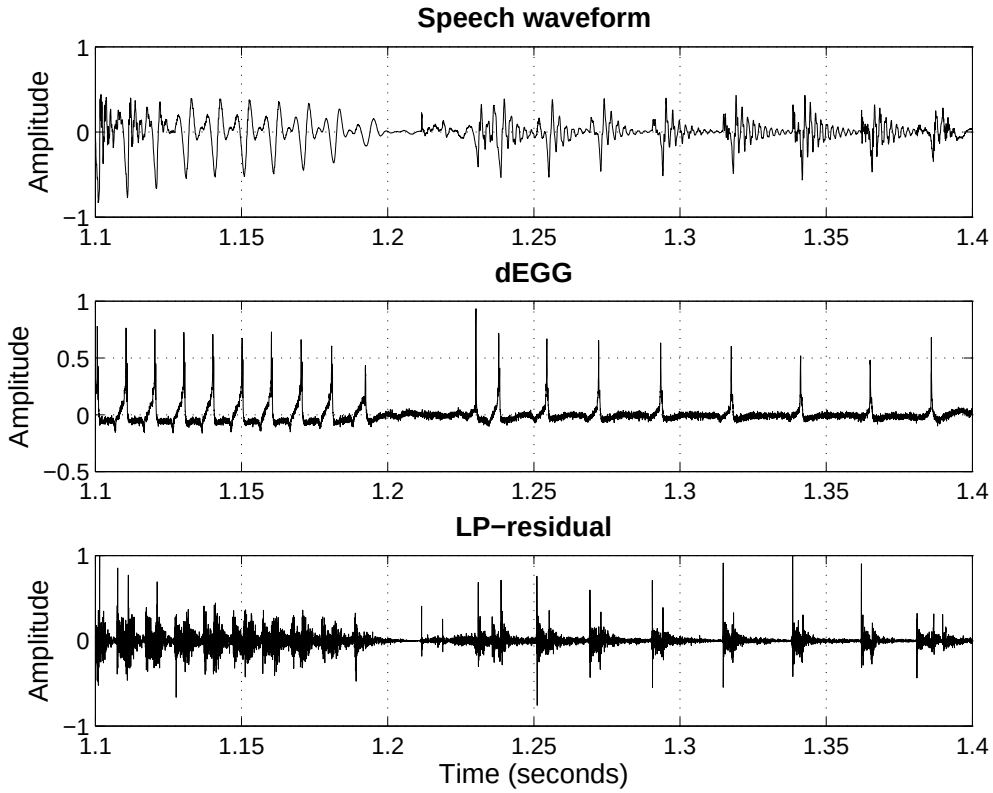


Figure 2.7: Speech waveform (top panel), derivative EGG (middle panel) and LP-residual (bottom panel) for a sentence containing creaky voice from around 1.22 seconds.

frequency.

These distinctive acoustic characteristics can cause problems for standard speech analysis methods (including f_0 tracking and spectral analysis). The very low f_0 values and, at times, irregular temporal patterning may not be properly handled by standard f_0 tracking algorithms. Furthermore, standard frame lengths (usually no longer than 32 ms) may be too short to capture two creaky glottal pulses lengths and, hence, will be unsuitable for obtaining strong periodicity information. As a result of this, creaky regions will be poorly modelled in most speech technology applications.

Some of these features are illustrated in Figure 2.7. Note that although there are paired LP-residual peaks in the creaky region (from around 1.22 seconds) the derivative EGG signal does not display secondary peaks.

Falsetto

van den Berg (1968) states that falsetto is a phonation type involving stronger amounts of adductive tension, medial compression and longitudinal tension compared with modal

voice. This was later corroborated by Esling (1984). Perhaps due to the increased longitudinal tension, the vocal folds become thin (Laver, 1980) and a reduced level of sub-glottal pressure is applied compared to modal voice (Van Riper and Irwin, 1958). Falsetto is typically associated with a high pitch resulting in short glottal pulses (van den Berg, 1968) and it is also reported to display a steeper spectral slope (Monsen and Engebretson, 1977).

2.1.3 The supralaryngeal vocal tract

The supralaryngeal vocal tract consists of the cavities above the larynx including the pharynx, the buccal cavity and the nasal cavity and is highly adaptable particularly through movement of the jaw, tongue and lips (Flanagan, 1972). The main articulatory organs in the vocal tract are: the tongue, the mandible, the velum and the lips. The fastest articulator is believed to be the tip of the tongue (Hudgins and Stetson, 1937).

Variation in position of the various articulators of the vocal tract have the important function of modifying the sound source created at the larynx (Hardcastle, 1976). For stop consonants a complete closure is created, which is followed by a build of air pressure and then a sudden release of air. The location of the occlusion determines where the turbulent noise is created which, in turn, affects the phonetic quality of the sound (Flanagan, 1972). Placement of the vocal tract articulators can also produce narrow constrictions at which point the airflow becomes turbulent giving rise to fricatives. Again the location of the constriction determines the phonetic quality of the consonant.

For nasal consonants the velum is opened and this is combined with complete closure of the vocal tract further forward, usually at the location of the velum, between the tongue and the alveolar ridge or at the lips (Flanagan, 1972).

2.2 Acoustic theory of speech production

The acoustic theory of speech production, described by Fant (1960), provides a theoretical framework on which much of the work on glottal source analysis is based. Fant (1960) describes a theory which allows for the functional separation of the speech production process into two main components: *source* and *filter*. This separation into two largely independent components allows the researcher to apply straightforward mechanisms from acoustic and electrical engineering theory. Since the initial publication in 1960, the theory has led to a large number of studies that have improved our understanding of the production of spoken language.

The theory facilitates a convenient close comparison of the engineering terms: *source*

and *filter* with the phonetic terms: *phonation* and *articulator*, respectively. From this, the source in voiced speech relates closely to the modulation of airflow by the vocal folds whereas the filter corresponds to what the listener perceives as the phonetic quality of the spoken utterance.

By applying the theory it is possible to decompose the speech waveform by removing the filter component (this is discussed further in Section 2.3.1). The residual signal is called the *voice source* or *glottal source* waveform. In the z-domain the source-filter theory can be stated as:

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

where the speech spectrum, $S(z)$, is the result of the spectral multiplication (or time domain convolution) of the glottal source, $G(z)$, the vocal tract filter, $V(z)$, and lip radiation, $L(z)$. Note that this representation of speech considers the excitation source for speech as beginning at the level of the larynx and does not explicitly model subglottal pressure from the lungs. The three components of the source-filter model of speech are a simplification of both the physiological mechanisms and the resulting acoustics. These three components are now described separately.

2.2.1 Glottal source

In the source-filter model the glottal source, $G(z)$, provides the excitation for the linear system. The glottal source is a model of the glottal flow, which is the airflow emanating from the lungs and modulated by the glottal area (Titze, 1994). The subglottal pressure from the lungs is not explicitly modelled and comes within $G(z)$. As is done elsewhere in the literature (see e.g., Degottex, 2010) the glottal source, $G(z)$, is treated as a signal description of the glottal flow which operates independently of changes in the vocal tract. The glottal source is mainly in two states: voiced or unvoiced. In its voiced stated $G(z)$ is a model of quasi-periodic vocal fold vibration, as well as any other laryngeal vibration which is used to excite the vocal tract system. The excitation of unvoiced speech is determined by the airflow becoming turbulent after passing through a narrow constriction, and this will be at different locations in the vocal tract depending on the phoneme (e.g., for /f/ the air will become turbulent at the labio-dental constriction (Flanagan, 1972).

2.2.2 Vocal tract filter

The vocal tract filter, $V(z)$, of the source-filter description models the effect of the supralaryngeal vocal tract. The vocal tract is sometimes treated as a uniform one-dimensional tube, open at one end, with lossless wave propagation. With this treatment, planar wave propagation can be assumed (Rabiner and Schafer, 1978) and it is also assumed that the vocal tract has the greatest cross dimension of less than a wavelength (i.e. for frequencies less than around 4 kHz, Flanagan, 1972).

This description of the vocal tract can be modelled using an all-pole model. Digitally this modelling of $V(z)$ involves p complex conjugate poles (Oppenheim and Schafer, 1989):

$$V(z) = \frac{1}{\prod_{k=1}^{p/2}(1 - c_k z^{-1})(1 - c_k^* z^{-1})} \tag{2.2}$$

where c_k and c_k^* are the complex conjugate pairs of the k^{th} formant.

Despite its simplification, the all-pole vocal tract model has seen wide application in speech processing (El-Jaroudi and Makhoul, 1991; Alku, 1992). However, it is widely believed that pole-zero pairs are required for proper modelling of nasalised sounds (Konvalinka and Matoušek, 1979; Xiaochuan et al., 2005). Recently developed speech processing methods have adopted a minimum phase modelling of the vocal tract filter (Bozkurt et al., 2005; Drugman et al., 2009a; Degottex et al., 2011a). This approach assumes that both poles and zeros exist within the unit circle. If the vocal tract is lossless the zeros will to lie on the unit circle (Lim and Lee, 1993). However, it can be hypothesised that the vocal tract losses will move the zero inside the unit circle (Degottex, 2010).

2.2.3 Radiation at the mouth and nostrils

The final component of the source-filter model is the radiation characteristic which occurs at the lips and/or nostrils. The radiating area around the lips and nostrils can be assumed to have a uniform velocity distribution and can be treated as a vibrating surface with all its parts moving in phase (Flanagan, 1972). Although more complex modelling can be used, the approach in speech analysis is often to model the radiation effects as a simple time derivative (Rabiner and Schafer, 1978; Wong et al., 1979; Markel and Gray, 1982):

$$L(z) = 1 - \alpha z^{-1} \tag{2.3}$$

where α is set here to 0.98. This corresponds to a single zero on the real-axis, just inside the unit circle. Although this filter is not strictly independent of sampling frequency, for

the range of sampling frequencies used in the current thesis (i.e. 10 kHz and 16 kHz) the difference is treated as being essentially negligible.

Using this model of the radiation effects one can considered the derivative glottal flow, $G'(z)$, as:

$$G'(z) = G(z)L(z) \tag{2.4}$$

and, hence, Eq. (2.1) can be reduced to:

$$S(z) = G'(z)V(z) \tag{2.5}$$

This reduction can be convenient as some glottal source models (e.g., the Liljencrants-Fant model, Fant et al., 1985a) model the glottal flow derivative.

2.2.4 Limitations of the theory

The separation of source and filter into two largely linear time-invariant (LTI) components is, however, an over-simplification of what actually occurs. Even if short segments of speech can be treated as time-invariant there are non-linear components involved. These non-linearities are due to source filter interaction effects and were the focus of research for part of the Ph. D. thesis by Lin (1990, Chap. 5). This acoustic interaction can result in a particular source pulse being affected by previous vocal tract oscillations (Fant and Lin, 1987). These interaction effects have three main consequences. *Skewing* of the glottal source pulse can occur (Fant and Lin, 1987; Rothenberg, 1981), which typically contributes to a stronger glottal excitation (Gobl, 2003). There may be a *ripple* effect imposed on the pulse from transglottal pressure variation as a result of formant oscillations from previous glottal pulses (Fant and Lin, 1987). Changes in formant frequencies and bandwidths within a single pulse duration can result in *damping* or *truncation* in the glottal pulse (Ananthapadmanabha, 1984; Fant et al., 1985b).

These interaction effects have been included in speech synthesis systems, but experiments have demonstrated only a slight difference in terms of the perception of the output (Nord et al., 1984, 1986; Lin, 1990).

The basis for the source-filter theory has been challenged most notably by Teager and Teager (1983, 1990). Teager and Teager (1990) argue that the production of speech is neither linear nor passive and that assuming as much would lead to erroneous and inconsistent conclusions.

Despite this, the evidence produced from speech research in the years since its inception

support the usefulness of characterising speech in this way (Gobl, 2003). Furthermore, many functional speech technology applications have been designed by exploiting the theory (see e.g., Cabral et al., 2011b; Raitio et al., 2011; Degottex et al., 2011a).

2.3 Characterising the glottal source

Many approaches to analysing the glottal source and voice quality involve exploiting the source-filter theory by attempting to separate the vocal tract filter effects from the speech signal and parameterising the residual (i.e. the glottal source estimate). This section reviews the most commonly use methods for source-filter decomposition and glottal source parameterisation.

2.3.1 Glottal inverse filtering

Glottal inverse filtering is the process of estimating the glottal source by deconvolution of an estimated vocal tract model from the speech signal. Modern glottal inverse filtering methods generally fall into three main categories: closed-phase methods, iterative methods and phase based methods. Note that three of the inverse filtering methods described here are evaluated in Chapter 6. Despite the attention the problem of glottal inverse filtering has received from the literature it is still believed that a fully functional automatic glottal inverse filtering method is yet to be developed (Walker and Murphy, 2007). In the discussion below a brief comment is given on the limitations of the various approaches. It may become apparent to the reader that despite a concerted effort from the research community involving a range of signal processing methods, glottal inverse filtering is yet to solved to a satisfactory level. As a consequence there may frequently be gross errors in the glottal source estimate which limit the effectiveness of subsequent parameterisation.

Closed-phase inverse filtering

Initial papers by Strube (1974) and Wong et al. (1979) looked to exploit the so-called 'closed-phase' region within each glottal pulse cycle as a means of deriving the vocal tract transfer function. During the glottal open-phase, where the vocal folds are apart, there is a certain amount of interaction between the glottal source signal and the vocal tract system with relatively strong damping of the vocal tract resonances (Yegnanarayana and Veldhuis, 1998). After the glottal closure instant (GCI, Naylor et al., 2007) and up until the moment of glottal opening, there is vocal tract resonance relatively free of effects from the glottal source. If this closed-phase region can be detected a covariance based Linear

Predictive Coding (LPC) approach (Rabiner and Schafer, 1978) can be used to determine an all-pole vocal tract model.

Several problems exist for the practical implementation of such a glottal inverse filtering approach. One is the difficulty in determining the closed-phase region. An algorithm was proposed in Alku et al. (2009) to improve the robustness to errors due to frame position by incorporating constraints in the model optimisation. Also, recent algorithms have been developed to improve the detection of the frame position (see Thomas et al., 2012 and Drugman et al., 2012c). However localisation of the glottal opening instant (GOI) is still significantly less robust than that of the GCI. Furthermore, in phonation types like breathy voice the vocal folds may not fully close making the detection of functional GCIs and GOIs all the more difficult. Another problem is that for speech with high f_0 values the duration of the closed-phase may be insufficient for stable vocal tract model estimation. To overcome this, a multicycle covariance method has been proposed (Yegnanarayana and Veldhuis, 1998; Plumpe et al., 1999) which involves combining closed-phase regions from neighbouring glottal pulses. Despite the improvements offered by this approach the performance of closed-phase inverse filtering still tends to degrade for signals with high f_0 values.

A closed-phase inverse filtering (CPIF) algorithm can be implemented as follows: GCIs and GOIs are detected (e.g., using the algorithm described in Drugman et al., 2012c which demonstrated improved robustness in determining these locations). Using the detected closed-phase, as marked by the GCIs and GOIs, covariance LPC analysis is carried out using:

$$\left(\sum_{k=0}^{K-1} \mathbf{C}_k\right) \mathbf{a} = -\left(\sum_{k=0}^{K-1} \mathbf{c}_k\right) \tag{2.6}$$

where $\mathbf{C}_k$ is the $p \times p$ covariance matrix and $\mathbf{c}_k$ is the covariance vector of the k^{th} closed-phase region, $\mathbf{a}$ is the vector (length p) of the prediction coefficients and K is the number of consecutive closed-phase regions used (Yegnanarayana and Veldhuis, 1998). In the implementation used in this thesis, K is set to 3 (i.e. the present closed-phase and the two adjacent ones) and the commonly used prediction order $p = fs/1000 + 2$ was used.

Iterative and adaptive inverse filtering

Several iterative algorithms have also been proposed for automatic inverse filtering. Recently a method was described based on the ARX (Auto-Regressive eXogenous) LF model for simultaneous estimates of glottal source and vocal tract models using iterative opti-

misation. A commonly used method (see e.g., Cabral et al., 2011b; Raitio et al., 2011) is the Iterative and Adaptive Inverse Filtering (IAIF) which is also included in this thesis. A block diagram of the algorithm is shown in Figure 2.8. In the present work the algorithm is applied to GCI centred frames, of twice the local glottal period in duration. The method works by successive vocal tract all-pole model estimation following the removal of the estimated glottal source contribution modelled with a prediction order which increases at each iteration. Originally prediction coefficients were determined by LPC analysis (Alku, 1992). A subsequent study, however, replaced LPC with discrete all-pole (DAP) modelling (El-Jaroudi and Makhoul, 1991), which involves the use of the Itakura-Saito distance measure (Itakura and Saito, 1968), in order to produce a better modelling of the spectral envelope for higher-pitch voices (Alku and Vilkman, 1994).

Despite the usefulness of the IAIF it can, nevertheless, output significant errors in the estimation of the glottal source signal. In particular, for vowels with a low first formant frequency (e.g., /i/) the interaction between the glottal formant and the first formant can lead to incomplete resonance cancellation in the outputted waveform (Alku, 1992). As a result any analysis of the estimated glottal source waveform will be negatively affected.

Mixed-phase decomposition

A final category of glottal inverse filtering involves decomposing the speech signal based on maximum and minimum phase components. In Doval et al. (2003), the authors demonstrate how the glottal source can be modelled as a combined causal-anticausal linear filter, where a pair of poles lying outside the unit circle (i.e. anticausal) correspond to the glottal formant and where a single pole lying inside the unit circle (i.e. causal) corresponds to the spectral tilt. This initial modelling was then used as the inspiration for the all-zero representation of the speech signal described in Bozkurt et al. (2005). Bozkurt et al. (2005) proposed a method called Zeros of the Z-Transform (ZZT), which utilises the unit-circle for separating zeros inside the unit circle (corresponding to vocal tract contribution and return phase of the glottal source) from zeros outside the unit-circle (corresponding to the open phase of the glottal source). A computationally more efficient method for separating these causal and anticausal components, utilising the complex cepstrum, was described in Drugman et al. (2009a). The complex cepstrum, $\hat{s}(n)$ can be derived from the speech signal, $s(n)$, using (Oppenheim and Schafer, 1989):

$$S(\omega) = \sum_{n=-\infty}^{\infty} s(n)e^{-j\omega t} \tag{2.7}$$

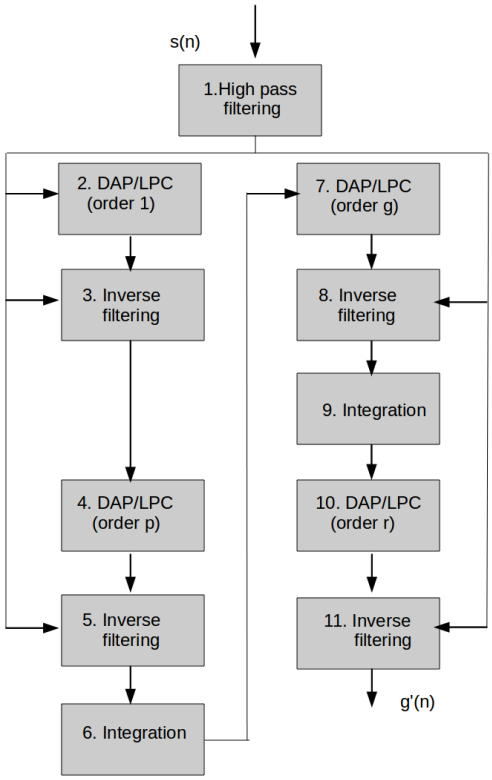


Figure 2.8: Block diagram of the Iterative Adaptive Inverse filtering (IAIF) method. Inputted is the speech signal, $s(n)$, with an estimate of the glottal source derivative signal, $g'(n)$, as the output. Different prediction orders are used: $p = fs/1000 + 2$, $g = 4$ and $r = p$

$$log[S(\omega)] = log(|S(\omega)|) + j\angle S(\omega) \tag{2.8}$$

$$\hat{s}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} log[S(\omega)]e^{j\omega}d\omega \tag{2.9}$$

Deriving the estimation of the glottal source (open phase) is then carried out by simply retaining the sample values below 0 quefrency.

The proper window position and type has been shown to be critical for suitable decomposition and, hence, the settings suggested in Drugman et al. (2009a) are used. This involves using a Blackman window, centred on a GCI and of two glottal periods in duration. This method was demonstrated to perform strongly compared with the state-of-the-art in a previous comparative evaluation (Drugman et al., 2011).

The main drawback of this approach is that there is no separation of the glottal return phase from the vocal tract filter, as both components are believed to be minimum phase.

For certain applications, e.g., parametric speech synthesis and voice modification, such a separation is essential. Furthermore, certain phonation types (e.g., breathy voice) exhibit glottal closing characteristics which can affect the robustness of GCI detection. As the mixed phase decomposition critically relies on proper window positioning which is based on the GCI, estimated glottal waveforms may be severely affected.

2.3.2 Glottal source models

The following is a brief description of the more commonly used models of the glottal source signal which are used in the literature.

Rosenberg model

The Rosenberg glottal model (model B in Rosenberg, 1971) is calculated using:

$$g_{ROS}(t) = \begin{cases} at^2 - bt^3 & \text{if} \quad 0 < t < t_e = t_c \\ 0 & \text{if} \quad t_c < t < T_0 \end{cases} \tag{2.10}$$

where t_e is the timepoint of the main excitation, t_c marks the beginning of the closed phase and T_0 is the glottal period duration. The parameters a and b control the duration of the open phase amplitude of voicing.

An example Rosenberg glottal model pulse is shown Figure 2.9 with the a parameter set to be equivalent to an open quotient (OQ) of 0.6, and with and F0 of 100 Hz. Note that this glottal model has an abrupt return to 0 following the main excitation. OQ is defined here as the duration of the open phase normalised to the glottal period, i.e. $\frac{t_e}{T_0}$.

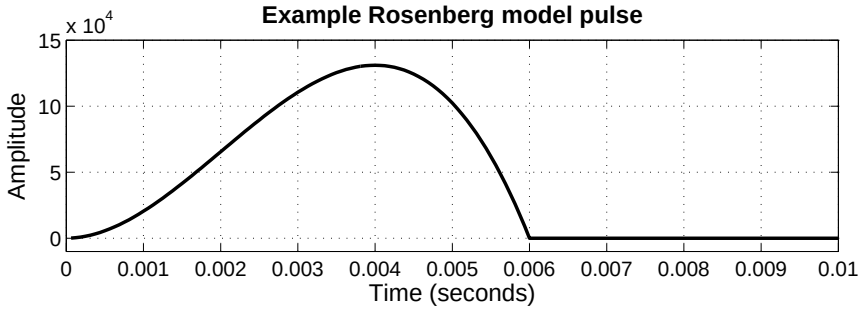


Figure 2.9: Example Rosenberg glottal model pulse.

KLGLOTT88 model

The formant synthesiser KLSYN88 (Klatt and Klatt, 1990) uses the glottal source model KLGOTT88[3]. The model is a third-order polynomial which can be smoothed using a low-pass filter. The KLGOTT88 is essentially a low-pass filtered version of the Rosenberg model, where the TL controls the spectral tilt of the model by setting the cut-off of a first order low pass filter. In the KLSYN88 synthesiser further parameters control other aspects of the glottal source model: AH; for determining the strength of aspiration noise, DI; for setting the amount of diplophonic irregularity of pulse durations and FL; for causing fluctuations in adjacent glottal pulse lengths.

LF model

The LF model (Fant et al., 1985a) is five parameter (including f_0 and assuming $t_c = T_0$) model of differentiated glottal flow that evolved from previous models developed by the same authors (see Fant, 1979 and the L-model also described in Fant et al., 1985a). Besides f_0, the LF model can be derived from parameters which consist of three time points: t_p, t_e and t_a as well as one amplitude value, EE (see Figure 2.10[4]). The model is made up of two components, the open-phase and the return-phase, and is calculated using:

$$g'_{LF}(t) = \begin{cases} E_0 e^{\alpha t} sin\omega_g t & \text{for} \quad t_o \leq t \leq t_e \quad \text{open-phase} \\ \frac{-EE}{\epsilon T_a}(e^{-\epsilon(t-t_e)} - e^{-\epsilon T_b}) & \text{for} \quad t_e < t < t_c \quad \text{return-phase} \end{cases} \tag{2.11}$$

where ω_g is $\frac{\pi}{T_p}$, $T_b = t_c - t_e$, α and E_0 are required to achieve area-balance (absolute area of the two segments is always the same) and ϵ is derived iteratively using:

$$\epsilon = \frac{1}{T_a} \cdot \left(1 - e^{-\epsilon T_b}\right) \tag{2.12}$$

Full details for solving α and E_0 are given in Gobl (2003) and Fant et al. (1985a). Both optimisation parts, i.e. for solving ϵ, and α and E_0 are done using the Newton-Raphson method.

The shape of the LF model pulse can also be characterised with the three so-called R-parameters: Rg, which is the frequency of the glottal formant normalised to f_0, is calculated using:

[3]Note that there are two other glottal models which can also be used in KLSYN88

[4]Note that for the LF model the glottal flow will rest exactly on the zero line. For natural glottal flow pulses, and in particular in laxer or breathier speech, there may be a strong DC component resulting in the glottal flow pulses being elevated off the zero axis.

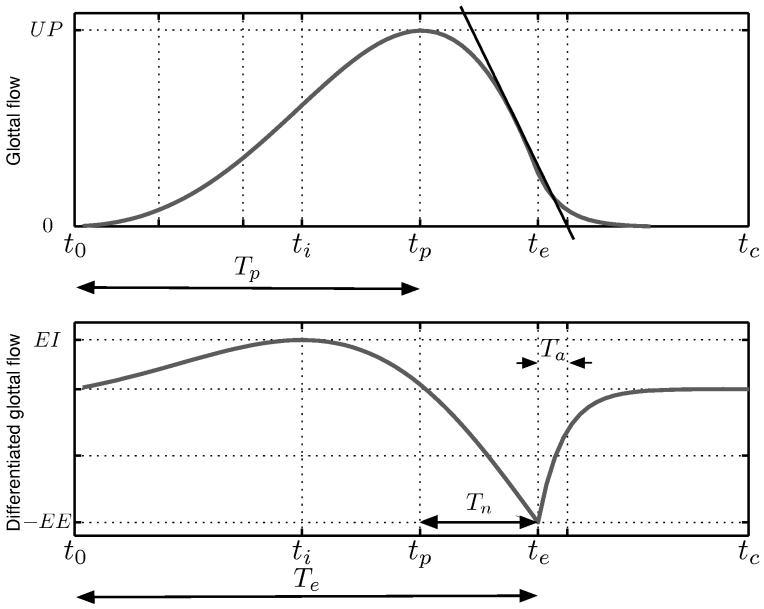


Figure 2.10: Example LF model pulse of glottal flow (top panel) and differentiated glottal flow (bottom panel).

$$Rg = \frac{T_0}{2T_p} \tag{2.13}$$

where T_p is the duration from glottal opening to the peak amplitude (Figure 2.10).

Rk is a measure of glottal skew and is the inverse of commonly used speed quotient. It is calculated as:

$$Rk = \frac{t_e - t_p}{t_p} \tag{2.14}$$

where t_e is the time point of the main excitation. Note that another asymmetry parameter also used in the literature is $\alpha_m = \frac{Rk}{Rk+1}$ (Henrich et al., 2001; Doval et al., 2006).

Ra defines the return phase and affects the level of attenuation of the higher frequencies in the spectrum. *Ra* is related to the extent of 'dynamic leakage', which is the residual airflow during the return phase (Gobl, 1988). *Ra* is derived with:

$$Ra = \frac{T_a}{T_0} \tag{2.15}$$

where T_a is the effective duration of the return phase.

A further *R*-parameter, *Rd*, was developed to provide a single parameter which captures most of the covariation of the LF model parameters (Fant et al., 1995). *Rd* is derived using:

$$Rd = 1000 \cdot \left(\frac{UP}{EE}\right) \cdot \left(\frac{f_0}{110}\right) \tag{2.16}$$

where UP is the peak amplitude of glottal flow and EE is the negative amplitude of the main excitation (differentiated glottal flow). The other R-parameters (Rg, Rk, Ra) can be predicted from Rd, following the regression analysis described in Fant et al. (1995), using the equations:

$$Ra_p = (-1 + 4.8Rd)/100 \tag{2.17}$$

$$Rk_p = (22.4 + 11.8Rd)/100 \tag{2.18}$$

$$Rg_p = \frac{Rk_p}{4(\frac{0.11Rd}{0.5+1.2Rk_p} - Ra_p)} \tag{2.19}$$

Note these equations are expected to hold for Rd values below an upper bound of around 3. For Rd values exceeding 2.7, Fant et al. (1994) suggest an amendment to these prediction equations. However, for Rd values above this level, the generated LF model has already begun to approximate a sinusoid.

The Open Quotient (OQ) parameter (originally proposed in Timcke et al., 1958), which is thought to be useful for discriminating breathy and tense voice (Henrich et al., 2001; Hanson et al., 2001), can also be derived from an LF model setting:

$$OQ = \frac{T_e}{T_0} = \frac{1 + Rk}{2Rg} \tag{2.20}$$

R++ model

Presented as an efficient alternative to the LF model and derived from the previous Rothenberg model (Rosenberg, 1971) the R++ model was designed in an attempt to match the flexibility of the LF model while at the same time improving computational efficiency (Veldhuis, 1998). The R++ model is derived using the same parameters as used for deriving the LF model. Perception tests that were carried out in Veldhuis (1998) suggest that there was little perceptual difference between synthetic speech generated using the LF model and the R++ model. Although the model has been used in some studies (e.g., Doval et al., 2006) the LF model tends to be more commonly used.

Causal anti-causal linear model

The causal anti-causal linear glottal flow model (CALM) is measured in the frequency domain (Doval et al., 2003). The authors believe that spectral based glottal source modelling has a number of advantages over time based modelling, e.g., that voice quality is better described by spectral parameters. Spectral tilt of the glottal source spectrum relates to the causal part of the CALM model. Doval et al. (2003) state that this can effectively characterise voice qualities in terms of loudness and weakness. The glottal formant corresponds to the anti-causal component of the model and this can allow characterisation of voice qualities in the tense to lax dimension.

Other glottal models

A range of other glottal source models have been proposed in the literature which have not been covered in this section (e.g., Ananthapadmanabha, 1984; Titze, 1984; Allen and Strong, 1985; Fujisaki and Ljungqvist, 1986; Shue and Alwan, 2010).

2.3.3 Glottal source parameterisation

A range of approaches have been proposed in the literature for parameterisation of the estimated glottal source signal. They are typically either direct measurements from the glottal source signal or are measurements derived from parametric models fit to the individual glottal source pulses (Strik, 1998). The different parameterisation methods can often suffer from a lack of robustness and some comments on this are given in the discussion below.

Direct measures

Time domain measurements of the glottal source are typically derived as quotients of the glottal period. They can provide useful information to do with important physiological timepoints in the glottal cycle. For instance, the open quotient (OQ) characterises the relative duration of the glottal open phase. Another measurement, the closing quotient (ClQ), is used as a relative measure of the glottal closing phase. However, the localisation of these timepoints is known to be problematic (Alku et al., 2002) and as a result parameters derived from amplitude based measurements that correlate with time domain quotients have been shown to improve robustness. The normalised amplitude quotient (NAQ, Alku et al., 2002), for instance, is calculated with:

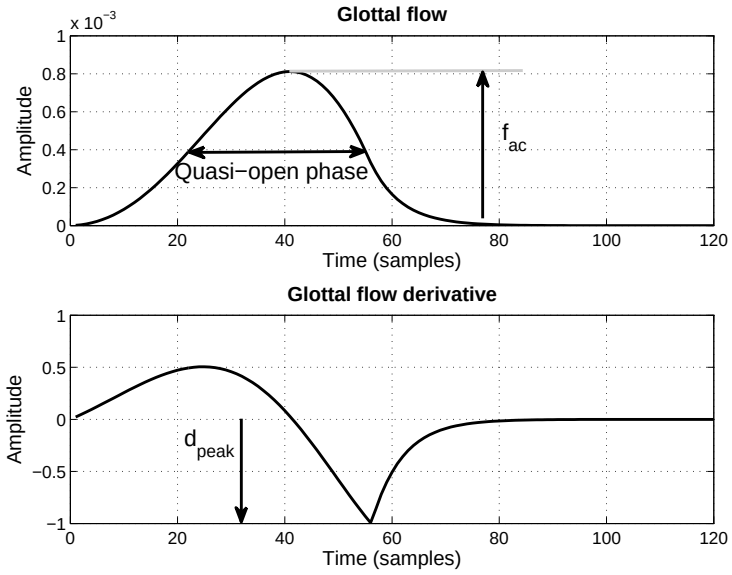


Figure 2.11: Glottal flow (top) and glottal flow derivative pulse (bottom) with the measurements required for deriving NAQ (f_{ac} and d_{peak}) and QOQ (quasi-open phase) highlighted.

$$NAQ = \frac{f_{ac}}{d_{peak} \cdot T_0} \tag{2.21}$$

where f_{ac} is the maximum amplitude of the glottal flow pulse and d_{peak} is the amplitude of the maximum negative amplitude of the glottal derivative pulse (see Figure 2.11). It has been shown to be a strong correlate of ClQ and has been practically used for analysing voice quality in non-ideal recording conditions (see e.g., Campbell and Mokhtari, 2003). However, some of the findings in (Gobl and Ní Chasaide, 2003a) suggest that the ability of NAQ to separate breathy and tense voice is reduced when there is high variability in f_0.

The quasi-open quotient (QOQ, Hacki, 1989) was developed as more robust measure relating to the standard open quotient. It is calculated by detecting the peak in the glottal flow and finding the time points previous to and following this point that descend below 50 % of the peak amplitude (see Figure 2.11). The duration between these time locations is used as a 'quasi-open phase' and is divided by the local glottal period in order to derive QOQ.

A number of parameters have also been developed which are derived from measurements in the frequency domain. The difference between the first two harmonics (H1-H2) in the narrowband amplitude spectrum of the differentiated glottal source signal is one such parameter. Another spectral parameter is the harmonic richness factor (HRF, Childers and Lee, 1991) which is measured as the sum of harmonic amplitudes above the first har-

monic divided by the amplitude of the fundamental. A further parameter, the parabolic spectral parameter (PSP, Alku et al., 1997), has been proposed for modelling the frequency domain characteristics of the glottal source. PSP is derived by fitting a parabola to the lower frequencies in the glottal source spectrum. Note, however, that the effectiveness of these measurements can be negatively affected when there is incomplete cancellation of lower formant resonance.

The above parameters rely on reliable estimation of the glottal source waveform. As glottal inverse filtering is known to lack robustness in certain types of speech regions (see Section 2.3.1) these parameters will often be negatively affected. Moreover, a circular problem arises as the very purpose of these parameters is to characterise different phonation types (e.g., breathy voice) and it is for these phonation types where their effectiveness may be most reduced.

Model fitting

Given an estimate of the differentiated glottal source signal, $g'(n)$, another way of characterising the salient aspects of the signal is to fit a model to the individual glottal pulses. A standard time domain based LF model fitting algorithm was described in Strik et al. (1993); Strik (1998). The algorithm is carried out as follows.

The method involves first finding an initial set of LF model-based parameters which are then refined during an optimisation procedure. In order to avoid the negative effect of high-frequency components on the fitting, a low-pass filtering operation is first applied. This is done by convolving the differentiated glottal source signal with an 11-point Blackman window, which has the characteristic of having a ripple-free impulse response (Strik, 1998). Fitting to each differentiated glottal source pulse initially starts from a given GCI (which can be detected, for instance, using the algorithm described in Chapter 4). The time point t_e (see Figure 2.10) can be obtained by searching nearby the GCI location for the maximum negative amplitude. This amplitude is chosen as EE. Then a search is done to the left of t_e for the first zero-crossing. This point is assigned as t_p. The point of glottal opening, t_o, is obtained by continuing the search left until pulse descends to below a small threshold. In this thesis the threshold used is 0.1 times the maximum positive amplitude of the present voice source pulse. In Strik et al. (1993) the authors suggest using an FFT based method to obtain an initial t_a value. However, the author's experience of using this approach is that it frequently gives unsuitable t_a values. Instead an initial Ra value of 0.02 is fixed, and t_a is calculated from this. A similar approach is used in the SKY voice source analysis software (Kreiman et al., 2006).

Initial parameters are then refined using a two-step optimisation approach. During each iteration of the optimisation methods LF model pulses are generated and an error value is measured between the generated pulse and the present voice source pulse. However, as the low pass filtering of the voice source signal affects the pulse shape, the same low pass filtering is applied to each generated LF model pulse. A root-mean-squared (RMS) error function is used in both optimisation steps. The first step involves the use of a simplex based method (Nelder and Mead, 1965) which is believed to be insensitive to initial fitting errors. Finally, a steepest descent algorithm is applied to further refine the LF model fit. In both steps all model parameters, EE, t_o, t_p, t_e and t_a are free to vary.

Similar to the direct measures, model fitting methods will also be negatively affected by errors in glottal inverse filtering. Furthermore, the method described above can also produce discontinuities in parameter trajectories even in relatively stable regions in the speech signal. This is often due to the difficult in consistently located the point of glottal opening (Alku et al., 2002).

An alternative method for LF model fitting can be carried out using amplitude based methods (see initial work on this in Kane and Gobl, 2009). LF model-based parameters can be derived using the equations proposed in Gobl and Ní Chasaide (2003a). First, the amplitude of the main excitation EE, the maximum amplitude of the glottal pulse (UP) and of the glottal derivative pulse (EI) need to be measured for each glottal cycle. The amplitude based estimation of Rg can then be calculated:

$$Rg_a = \left(\frac{1}{\pi}\right) \cdot \left(\frac{EI}{UP}\right) / f_0 \tag{2.22}$$

and Rk can be estimated with:

$$Rk_a = \left(\frac{2}{\pi}\right) \cdot \left(\frac{EI}{EE}\right) \tag{2.23}$$

There then remains just a single shape parameter, Ra, required to characterise the LF pulse. An initial Ra value can be estimated from Rd (derived from amplitude values using Eq. 2.17). This Ra estimate can then be optimised in the frequency domain by minimising the following error criterion:

$$e = \frac{1}{N} \sum_{f=0}^{N} \frac{P(f)}{\hat{P}(f)} - ln\frac{P(f)}{\hat{P}(f)} - 1 \tag{2.24}$$

which is the discrete Itakura-Saito distance measure between the power spectrum of the windowed differentiated glottal source frame $P(f)$ and the synthetic differentiated glottal

source frame $\hat{P}(f)$, where f is frequency in Hertz and N is the maximum frequency considered. In this thesis N is fixed at 3 kHz. Note that the synthetic differenitated glottal source frame is generated using the already measured LF model-based parameters and the present estimated Ra value. This procedure for LF model fitting will be called Amp-LF in this thesis.

Glottal model estimation by phase minimisation

A recently developed algorithm allows derivation of the Rd parameter without prior inverse filtering and without explicit model fitting. The method is described in Degottex et al. (2011b). It involves estimation of the Rd parameter of the LF model by minimising a phase-based criterion and has the advantage of being calculated independently of the precise position of the glottal pulse and of the strength of the main excitation, EE. A brief summary of the algorithm is given.

The algorithm considers the so-called mean squared phase difference (MSPD). One can utilise the outcome of minimising:

$$MSPD^2(\theta, N) = \frac{1}{N}\sum_{k=1}^{N}\left(\Delta^{-1}\Delta^2 \angle R_k^\theta\right) \tag{2.25}$$

where $N = |f_{lim}/f_0|$, k is the harmonic index and θ is the shape parameter of the glottal model (i.e. Rd). The computation of this objective function involves applying the second order phase difference (Δ^2) and the anti-difference operator (Δ^{-1}) to the *convolutive residual*, R_k^θ. R_k^θ is the deconvolution of the given speech spectrum, S_k, by the speech model. The speech model is the combination of the glottal pulse shape, G_k^θ, with the minimum phase vocal tract model, $\varepsilon_{\text{-}}(S_k/G_k^\theta \cdot jk)$ (where $\varepsilon_{\text{-}}(.)$ is the minimum phase version of a given spectrum).

$MSPD^2(\theta, N)$ is minimised with respect to θ (i.e. Rd) using the algorithm described in Brent (1973) which has the advantage of not requiring initial parameter estimates.

A more recent study (Huber et al., 2012) has proposed variants of this phase minimisation criterion as well as further development of the regression analysis used with the Rd parameter, originally described in Fant et al. (1994, 1995).

One striking advantage of this method is that it avoids glottal inverse filtering. However, as the method depends on phase characteristics of the signal it is likely that the effects of phase distortion, which could result from the recording equipment, would reduce the robustness of the analysis. Such effects of phase distortion were recently discussed in O' Cinnéide (2012)

2.3.4 Simultaneous source-filter parameterisation

The parameters of the ARX-LF model, which applies an Auto-Regressive vocal tract model excited by the combination of an eXogenous signal with an LF model pulse (Vincent et al., 2007), are often solved simultaneously for a given pitch-synchronous speech frame.

A time-domain approach to solving the parameters of the ARX-LF model was described in Hui-Ling (2002) and del Pozo (2008). The algorithm initially uses the parameters of the KLGLOTT88 model and derives these parameters along with an all-pole vocal tract model parameters by applying a convex optimisation method to minimise a time domain squared error of the residual signal. Dynamic programming is used to determine the optimal path of the vocal tract and KLGLOTT88 parameters and finally an inverse filtering operation is carried out using the derived vocal tract model and the glottal source estimate is re-parameterised this time using the LF model.

2.3.5 Manually optimised glottal source analysis

Another approach to inverse filtering and glottal source parameterisation is to combine automatic methods with manual optimisation. A comprehensive procedure for carrying out this process was outlined in Gobl and Ní Chasaide (2010, 1999a). For the inverse filtering part, this procedure first involves applying automatic closed-phase inverse filtering. The inverse filtering is then optimised manually by allowing the user to adjust formant frequencies and bandwidths to achieve maximum formant cancellation. The user is guided through the use of time and frequency domain displays and seeks to cancel resonance in resulting spectrum and also reduce any time-domain ringing, particularly in the closed phase.

Similarly, for the parameterisation part, model matching (using the LF-model) is done automatically. Again the user optimises the fit (by modifying time and amplitude model parameters) using time and frequency domain displays. In order to optimise the fit the user typically attempts to ensure close fitting to the region around the main excitation. The user also pays particular attention to the consistent marking of the point of glottal opening. Often the estimated glottal derivative pulse does not quite cross the zero-line as can be seen in Figure 2.12. Using the actual zero-crossing (indicated by the negative red stem) would clearly result in over estimation of the glottal open phase and, hence, the manual user may at times have to visually extrapolate from the curve of the open phase lobe in order to more accurately and consistently mark the point of glottal opening.

The obvious drawback of this approach is that it is extremely time-consuming, as each individual glottal pulse needs to be analysed. Another drawback is that the user needs to

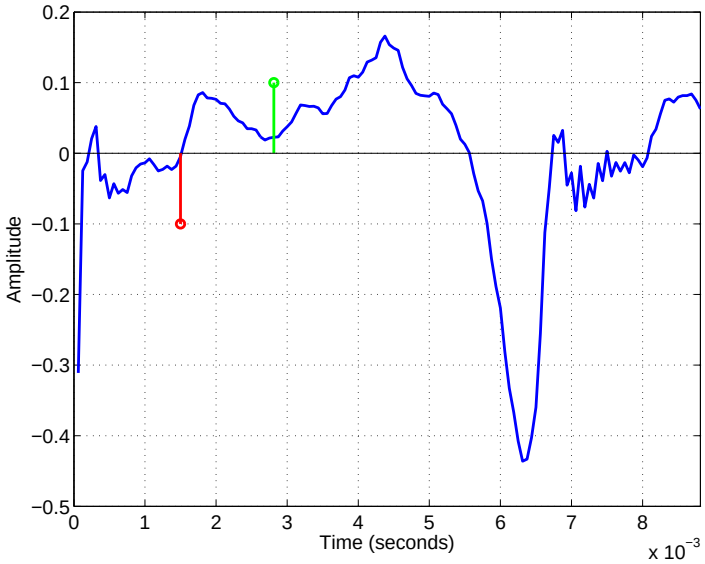


Figure 2.12: Differentiated glottal pulse illustrating the potential for inconsistency in marking of the point of glottal opening. The positive green stem shows the marking of glottal opening by a user of the manually-optimised approach, with the negative red stem showing its estimation by an automatic algorithm

be experienced in this type of analysis and that there is a danger of subjectivity. Despite this, one can obtain very precise glottal source measurements using this approach which may otherwise not be obtainable using purely automatic approaches.

2.4 Summary

This chapter provided an overview of the production and acoustics of speech, with particular attention paid to the laryngeal activity relevant for producing different phonation types. The chapter also discussed different approaches to doing acoustic modelling of speech based on Fant's source-filter theory. A survey of the different approaches to estimating the glottal source by glottal inverse filtering was given, with a subsequent review of the different approaches used to characterise the glottal source signal.

Chapter 3

The glottal source in spoken communication and speech technology

3.1 The role of the glottal source in speech communication

As humans we exploit the different modalities that are available to us in order to communicate with each other. Changes in the articulation of the vocal tract are predominantly used for achieving phonetic contrast. However, even from a very young age infants are sensitive to variation in a parent's voice quality long before they have any real grasp of the linguistic content of the utterances (Marwick et al., 1984; Mackenzie Beck, 2005). As phonation type is a main component for much of the variation in voice quality, we develop a strong command of our vocal folds. We use our voice to carry the linguistic content of our spoken utterances and colour them with aspects of our communicative intention, our affective states, etc.

The following section provides a broad illustration of how variation in the voice is exploited by speakers. Note that this section does not seek to provide a comprehensive survey of the area but rather serves as an illustration of the importance of the voice in different aspects of spoken communication.

3.1.1 Voice quality

Gross changes in phonation type can bring about clearly perceivable differences in voice quality and these changes are utilised by speakers for both linguistic and so-called paralinguistic functions in speech.

Linguistic function of voice quality

Some languages use short-term changes in voice quality for pronouncing certain linguistic units (Laver and Trudgill, 1979). Below are some examples of how phonemes are differentiated using voice quality variation in certain languages. The set of examples is not intended to be anywhere near exhaustive but simply serves to provide an illustration of how voice quality can also be used to affect the propositional content in some languages.

In Gujarati, speakers use modal and breathy voice to contrast some vowels e.g., the words **/baɾ/** ('twelve') and **/b̤aɾ/** ('outside'), are contrasted solely on the presence or absence of breathiness in the vowel (Ladefoged and Maddieson, 1996, p. 315). In the Central American language Jalapa Mazatec both breathy and creaky voice are used in the production of certain vowels (Silverman et al., 1995). For instance, **/j̤æ̤/** ('a boil') and **/g̤i̤/** ('he went') involve the use of breathiness in the vowel. Creaky voice is also used, for example, in **/s̰ḭ/** ('holiday') and **/tʃṵ/** (blouse'). Also, it is common in some American Indian languages to use creaky voice for contrast between some vowels and nasals (Gordon and Ladefoged, 2001). Harsh voice has also been noted in the phonetic system of !Xóõ (Traill, 1986).

Historically, the Khmer (Cambodian) language exploited the difference between breathy and clear or modal voice for phonetic contrast. One dialect of Khmer spoken in Chanthaburi Province in the east of Thailand maintains the use of these phonation types to differentiate some vowels (Wayland and Jongman, 2003). For instance, the minimal pairs **/t̤i̤ɛn/** ('candle') and **/tiɛn/** ('to blame') as well as **/k̤a̤t/** ('he') and **/kat/** ('to cut') are differentiated based just on whether there is breathiness in the vowel.

In terms of voice quality use in tonal languages a 'growling' harsh tone has been reported in Zhenhai (Rose, 1989). The higher pitch *ventricular voice* (Laver, 1980) has also been observed in the tones of Bai (Esling and Harris, 2003).

Paralinguistic function of voice quality

The term paralinguistic is often used in the literature to refer to the communication of affect and other intentions that are beyond the normal scope of linguistics (Gobl, 2003). However, there is no obvious boundary separating the linguistic from the paralinguistic

and in fact the two are inextricably coupled when considering interactive, conversational speech. Nevertheless, as the term is used widely in the literature a separate discussion is given here on the paralinguistic function of voice quality. Paralinguistics is used broadly here to refer to the intentions, attitudes, emotions and affective states expressed by speakers (Ishi et al., 2008a).

Speakers in natural conversational settings regularly deviate from their habitual voice quality to signal attitude, mood and emotion (Gobl and Ní Chasaide, 2003b). It has been suggested that some of these aspects share universal characteristics (Brown and Levinson, 1987), though it is likely than many aspects are specific to language and culture (Ogarkova et al., 2009; Yanushevskaya et al., 2011).

Breathy voice has been generally observed in association with intimacy and familiarity (Laver, 1980). It has also been studied in relation to politeness in Japanese (Ito, 2004). Typically, sustained high pitch can be used when conveying politeness (Brown and Levinson, 1987), although this is more likely to be associated with femininity than politeness in Japanese. It is, hence, suggested by Ito (2004) that Japanese speakers, particularly males, use other vocal strategies (i.e. breathiness) in order to communicate politeness. Kasuya et al. (2000) and Fujimoto and Maekawa (2003) also reported that breathiness was involved in Japanese expression of disappointment.

Creaky voice is often reported to be used for signalling boredom or resignation (Cruttenden, 1986). It is also frequently used for signalling a range of paralinguistic information in Japanese (Sadanobu, 2003) and in Finnish (Silen et al., 2009). The use of creaky voice has been studied in relation to turn-taking (Ogden, 2001) and hesitations (Carlson et al., 2006) in interactive speech, as well in the context of various forms of expression and emotion (Gobl and Ní Chasaide, 2003b; Ishi et al., 2008a; Yanushevskaya et al., 2011).

Falsetto is a phonation type typically associated with high pitch and is exploited as an expressive tool in spoken conversation, sometimes associated with fear (Schroeder, 2001) as well for emphasising femininity (Podesva, 2007).

The use of *whisper* and *whispery voice* is frequently used to convey secrecy or confidentiality in English and many other languages (Laver, 1980; Fan and Hansen, 2013).

Some more general studies on the perception of voice quality report a mapping between lax-creaky voice and the affects: boredom, sadness and relaxedness (Gobl and Ní Chasaide, 2003b; Yanushevskaya et al., 2005).

Although the above discussion provided some specific examples of direct relationships between voice quality and the signalling of paralinguistic information, in everyday spoken conversation these relationships are more complex. Furthermore, many of the observations in the studies referred to above involved qualitative inspection of at times small volumes of

speech data. However, in order to quantitatively study the use of voice quality in natural settings, effective and robust analysis tools are required.

3.1.2 Glottal source dynamics in intonation and prosody

Although instances of non-modal voice quality involve sharp changes in phonation type, the glottal source is, in fact, continually varying in voiced speech (Gobl, 1988; Ní Chasaide et al., 2011). Recent studies (Yanushevskaya et al., 2010; Ní Chasaide et al., 2011) have investigated the variation of parameters describing the glottal source and have found some of these parameters to be important for the signalling of prominence or focus in spoken utterances. In fact in some instances certain glottal source parameters displayed peaks or dips more tightly aligned to the focused syllable than f_0. One specific finding was that the open quotient (OQ) parameter, in many instances produced a clear dip in the region of the focused syllable, suggesting a tenser phonation type. Vainio et al. (2010) also investigated the effect of expressions of focus on glottal source parameters and found that the Finnish speakers they analysed displayed normalised amplitude quotient (NAQ, Alku et al., 2002) values that suggested a laxer phonation type. Of course these studies do not necessarily contradict one another, as speakers of different languages may adopt varying vocal strategies for achieving prominence. These studies do, however, emphasise the importance of the glottal source for such communicative goals.

3.2 Applications and the impact on speech technology

In previous times when speech synthesis was mainly achieved through the use of formant or articulatory synthesis, the glottal source component played a critical role. This resulted in a large amount of research attention being paid to the estimation and modelling of the glottal source and its inclusion in speech synthesis (see e.g., Klatt, 1980; Holmes, 1983; Carlson et al., 1989; Mahshie and Gobl, 1999). However, with the advent of non-parametric concatenative speech synthesis and Hidden Markov Model (HMM) based statistical speech synthesis (which typically uses either an impulse train or white noise as the excitation source) the importance of glottal source analysis faded somewhat. Recently, however, there has been a resurgence in the exploitation of the glottal source in speech technology through the use of emerging and state-of-the-art developments and approaches.

This section provides an illustration of some recent speech technology developments

which incorporate glottal source and voice quality analysis.

3.2.1 Speech synthesis and voice modification

The attraction of having flexibility of voice characteristics of speech technology applications has existed for some time. Recently there have been some notable developments in this area.

The work described in Cabral et al. (2008, 2011b) demonstrated how a glottal source model (Fant et al., 1985a) could be incorporated into a modern HMM-based synthesis platform. The method involves estimation of the glottal source signal using a conventional inverse filtering method and subsequent glottal source parameterisation. The glottal model signal is then removed from the speech signal in the frequency domain through the use of glottal spectral separation (GSS, Cabral et al., 2008). The resulting spectral envelope is then parameterised using the STRAIGHT algorithm (Kawahara, 1997) which ensures high quality synthesis. At synthesis time the parameters of the glottal model can be varied in order to approximate an alternative voice quality.

Another state-of-the-art statistical parametric synthesis system was recently proposed (Raitio et al., 2008, 2011) which also exploits glottal inverse filtering. For this method, however, the glottal source estimate is not parameterised with a glottal source model but instead modification and concatenation of natural glottal source pulses is used to create an excitation signal at synthesis time. This excitation source can also be modified to change voice characteristics and has been shown to produce highly natural speech compared to previous state-of-the-art methods.

Glottal source modelling has also been utilised in recent voice modification methods. In the approach called *Separation of the Vocal-tract with the Liljencrants-Fant model plus Noise* (SVLN, Degottex et al., 2011a, 2012), the *Rd* parameter of the LF model is derived using the phase minimisation approach (see Chapter 2, Section 2.3.3 and Degottex et al., 2011b). The approach also includes parameterisation of the noise level, as well as it's temporal modulation, and a minimum-phase spectral envelope as the vocal tract filter. This has been shown to be particularly useful, compared to the state-of-the-art, for large pitch transposition and also facilitates alteration of the level of breathiness in the output. The SVLN method has also been utilised for improving the quality of statistical speech synthesis (Degottex et al., 2012).

A host of other voice modification methods involve glottal source modelling, often as part of the ARX-LF modelling of speech (see e.g., Vincent et al., 2005; Agiomyrgiannakis and Rosec, 2009; O' Cinnéide et al., 2011).

3.2.2 Separating speaking styles

Glottal source and voice quality parameters have also been shown to be useful for the purpose of separating speaking styles in corpora of expressive speech (Székely et al., 2011, 2012b). This sort of speech processing can help facilitate the development of speech synthesisers with different speaking styles for the one voice (Székely et al., 2012a). A study of political speeches found that voice quality parameters for discriminating breathy to tense voice correlated well with human annotations of speaking styles, compared to more conventional audio and video features (Scherer et al., 2012b). Furthermore, a recent study involving classification of different levels of vocal effort from expressive speech found that the inclusion of voice quality features (including the *Rd* estimation using the method in Degottex et al., 2011b) brought a significant improvement to the classification when combined with mel-cepstral coefficients (MFCCs) than when using MFCCs alone (Obin, 2012).

3.2.3 Emotion classification

The voice is intuitively a means by which people can express their affective states and not surprisingly measurements to do with the glottal source and voice quality have been found to be useful for the classification of emotions. Some notable work by Lugger and Yang (2007) investigated the contribution of the voice quality parameters described in Lugger et al. (2006) for classification of six emotions: anger, happiness, sadness, boredom, anxiety and neutral. The voice quality parameters used were the so-called glottal gradients (Lugger et al., 2006) which were developments of previous measurements described in Stevens and Hanson (1994). The study found that the voice quality parameters contributed additional information to the classification when combined with more commonly use prosodic parameters. A further study found improved classification of acted emotions when using glottal source parameters, including NAQ and OQ (Sun et al., 2009).

Several other studies have noted improvements in emotion classification by the inclusion of features describing the glottal source (Lliev et al., 2010; Tahon et al., 2012; Sun and Moore, 2012). More specifically, some studies have focused on the detection of depression and have emphasised the importance of glottal source features (Moore et al., 2003; Ozdas et al., 2004; Moore et al., 2008).

3.2.4 Other areas of speech technology

Aside from the main areas which have utilised glottal source parameters discussed above, other areas of speech technology have also benefitted from the inclusion of these parameters. For instance, a recent study (Zelinka et al., 2012) considered the effects of different levels of vocal effort (e.g., whisper, normal voice, shouting) on the performance of speech recognition of isolated words. The study presented a 50 % relative reduction in word error rate by normalising for vocal effort. Also, Zhang and Hansen (2007) investigated the effect of varying vocal effort from whisper to shouted on speaker identification and found that whisper caused the most significant deterioration in performance.

In Drugman and Dutoit (2010) the authors demonstrate how glottal signatures, derived from the Linear Prediction (LP) residual signal could be used for speaker recognition. Other work on speaker recognition observed improvements by including glottal model parameters and closed-phase information as input features to the classifier (Slyh et al., 2004). Furthermore, very recent research has demonstrated that more commonly used features of the glottal source have been shown to provide complementary information to standard spectral features for discriminating speakers (Félix Torres and Moore, 2012).

3.3 Summary

This chapter provided an illustration of the use of the glottal source in spoken communication. In particular changes in the phonation type, and consequently the glottal source, can be made to bring about changes in voice quality. This can be used for phonetic contrast in certain languages, as well as for various communicative and paralinguistic functions in interactive speech. The impact of glottal source characterisation on speech technology was also discussed, and it is clear that many areas of speech technology have recently benefitted from the inclusion of acoustic features to do with the glottal source and vocal tract filter separately. Nevertheless, there is a critical requirement for the improvement in the robustness of glottal source and voice quality measurements to increase the effectiveness of such features in speech technology.

3.4 Aims

The discussions in the present and previous chapter (i.e. Part 1 of the thesis) have provided the motivation for the set of aims of this thesis, which are now listed:

- To assess the effects of different phonation types on glottal closure instant detection and develop a new method to properly deal with these effects.
- To improve the robustness of automatic glottal source modelling.
- To develop novel methods for discriminating and detecting different commonly occurring voice qualities.
- To evaluate the newly developed algorithms against the state-of-the-art methods.
- To illustrate how these new algorithms may be exploited in functional applications.
- To construct a package of software containing code of the developed methods to encourage testing, usage and feedback.

Part II

Fine-grained analysis methods

Chapter 4

Glottal closure instant detection in a range of voice qualities

Summary

Recently developed speech technology platforms, such as statistical speech synthesis and voice transformation systems, facilitate the modification of voice characteristics. To fully exploit the potential of such platforms, speech analysis algorithms need to be able to handle the different acoustic characteristics of a variety of voice qualities. Glottal closure instant (GCI) detection is typically required in the analysis stages, and thus the importance of robust GCI algorithms is evident. The current study examines some important analysis signals relevant to GCI detection, for a range of phonation types. Furthermore, a new algorithm (called SE-VQ) is proposed which builds on an existing GCI algorithm to optimise the performance when analysing speech involving different phonation types. Results suggest improvements in the GCI detection rate for creaky voice due to a reduction in false positives. When there is a lack of prominent peaks in the Linear Prediction residual, as found for breathy and harsh voice, the results further indicate some enhancement of GCI identification accuracy for the proposed method.

4.1 Introduction

A required starting point for many speech and voice quality analysis methods is to obtain glottal closure instant (GCI) locations (Naylor et al., 2007). GCIs refer to the moments of most significant excitation that occur at the level of the vocal folds during each glottal period (Smits and Yegnanarayana, 1995). Knowledge of these locations can be used for applications like prosodic modification (Rao and Yegnanarayana, 2006), join optimisation in concatenative synthesis (Stylianou, 1999) and glottal inverse filtering (Drugman et al., 2009a). Due to the wide range of uses, GCI detection has received a considerable amount of research attention, with many studies focusing on the robustness of algorithms in degraded conditions (e.g., in the presence of noise, Drugman and Dutoit, 2009, distance from microphone, Guruprasad et al., 2007.)

During the last few years a range of speech technology platforms have been developed which facilitate the modification of voice characteristics. These platforms have come in the form of speech synthesis (Cabral et al., 2011b; Raitio et al., 2011) and speech modification systems (Agiomyrgiannakis and Rosec, 2009; O' Cinnéide et al., 2011). For instance, the speech synthesis system developed by Cabral et al. (2011b) uses a glottal source model (Fant et al., 1985a) as the excitation signal which can be modified by the user to approximate an alternative voice quality. The same glottal source model is used in the speech modelling methods described by Vincent et al. (2005) and O' Cinnéide et al. (2011). In both methods the shape of the glottal source pulses can be adapted in order to create a speech output with a laxer quality.

However, in order to fully exploit the potential of these platforms it is necessary to deploy speech analysis algorithms, which can cope with the wide range of voice quality variation often found in conversational speech. The importance of voice quality changes in speech communication was discussed in Chapter 3. Note that in this chapter the terminology in relation to voice quality and phonation types is consistent with that described in Chapter 2, Section 2.1.2.

Non-modal phonation types display varying glottal source characteristics, and this would likely impact on the performance of GCI algorithms. For example, creaky voice displays very different glottal characteristics than modal voice, with considerably longer pulses and at times irregular temporal patterning. If a GCI method could handle this difference then it may be possible to sufficiently model 'creakiness' in order to incorporate it in the speech output. Such a modification may be desirable for producing more 'conversational' or expressive synthetic speech. In fact, the present author has been engaged in work on the development of a vocoder which can give a natural rendering of creaky voice

for statistical parametric speech synthesis (Drugman et al., 2012b). However, for this a GCI algorithm which can handle creaky phonation is critical.

A recent study (Cabral et al., 2011a) demonstrated the effect that variation in phonation type had on GCI detection performance. Specifically, lax and breathy voices were shown to generally reduce the accuracy of many algorithms in terms of distance from estimated GCIs to the reference. Apart from this study and a paper which focused on the 'smoother glottal closure' found in breathy voice and voice offsets (Tuan and d'Alessandro, 199), research in this area has tended not to focus on development and evaluation of algorithms for non-modal phonation.

Hence, the current chapter presents, for a range of phonation types, an investigation of the characteristics associated with the closing of the glottis, in terms of different analysis signals typically used in GCI detection (Section 4.2). Furthermore, several widely used state-of-the-art GCI detection algorithms and the strategies employed in them are described (Section 4.3), as well as a newly developed algorithm which builds on a previously described method (Drugman et al., 2012c) modified in order to deal with some of the difficulties posed by non-modal phonation (Section 4.4). These algorithms are first evaluated on standard speech databases and also on speech produced by six speakers in a range of phonation types (Section 4.5). This work has recently been published in Kane and Gobl (2013)

4.2 Glottal closing characteristics of different phonation types

In this section the characteristics of a range of phonation types covered in the current study are described in terms of analysis signals typically used when considering GCI detection. In the descriptions observations are made based on the present data combined with previous evidence from the literature. As is done in Laver (1980) the phonation types are described in relation to modal voice, which is used as a reference. [1]

[1]For mainly pragmatic reasons, there is a slight deviation from Lavers terminology with regard to tense voice: Tense voice is here referred to as a phonation type, distinct from harsh voice. Note however, tense voice is generally thought of as involving elevated tension settings throughout the entire vocal system (Laver, 1980). At the laryngeal level, the increased tension may sometimes result in a phonation type which can be adequately described as harsh voice. However, compared to harsh voice, the increase in laryngeal tension in tense voice is typically less extreme, and may not give rise to the characteristic irregularities in vocal fold vibration associated with harsh voice (Laver, 1980). Hence, tense voice is here used to refer to a phonation type involving an increase in the tension settings compared to modal voice, but which does not display the irregular vocal fold vibration associated with harsh voice.

Two of the main analysis signals used when describing GCI detection are the first derivative of the electroglottographic signal (DEGG) and the linear prediction residual (LP-residual). The DEGG is the derivative of the EGG signal which is obtained through the use of a laryngograph. The device involves sending an electrical current between electrodes placed on either side of the larynx. The EGG signal is a measure of conductance which oscillates at the frequency of f_0 during voiced speech and which peaks when the glottis is closed. The DEGG signal is typically used for obtaining reference GCI values. The negative-going zero-crossings, following the peaks are typically used for obtaining the reference GCI.

The LP-residual is a commonly used analysis signal for localising glottal closure instants (Ananthapadmanabha and Yegnanarayana, 1979; Naylor et al., 2007; Drugman et al., 2012c). The signal is derived here by inverse filtering the (un-pre emphasised) speech signal using the coefficients obtained by autocorrelation linear prediction. In this work the order of the LPC analysis is determined by the sampling frequency, fs, according to $fs/1000+2$ (as was used in Drugman et al., 2012c) [2]. This order corresponds roughly to two coefficients to characterise each formant, for a typical male speaker, and two further coefficients to characterise the voice source contribution. The setting of this order, however, does not critically affect the GCI detection performance.

In the left column of Figure 4.1 the speech waveform, DEGG and LP-residual are displayed for a portion of a modal utterance produced by a female speaker. It can be seen in the bottom left panel that there are clear peaks in the LP-residual, some of which are aligned to the negative zero-crossings of the DEGG signal (middle panel). In comparison, the LP-residual for a breathy utterance (right column) is considerably noisier with less prominent peaks, despite the DEGG showing the phonation to be strongly periodic. Smoother closing characteristics have been previously reported (Tuan and d'Alessandro, 199) and perhaps arise from weaker laryngeal tension, compared with modal voice. This may partly account for the lack of impulsive peaks in the LP-residual.

Focusing on tense voice (left column of Figure 4.2) one can observe peaks in the LP-residual which are even more prominent than those in modal voice. Tense voice is often reported to display strong higher harmonics, compared to modal voice (Gobl and Ní Chasaide, 1992), which contributes to a more prominent peak in the LP-residual, corresponding to GCIs, because more harmonic components exceed the noise level in the higher

[2]This LPC order is chosen to obtain a spectral envelope which avoids fitting to the harmonics. Although it is widely known that LPC analysis becomes biased towards harmonics for high f_0 values (Villavicencio et al., 2006; Kay, 1988), there is no reported evidence, so far, of other spectral envelope estimation algorithms bringing improved GCI detection for higher pitched voices.

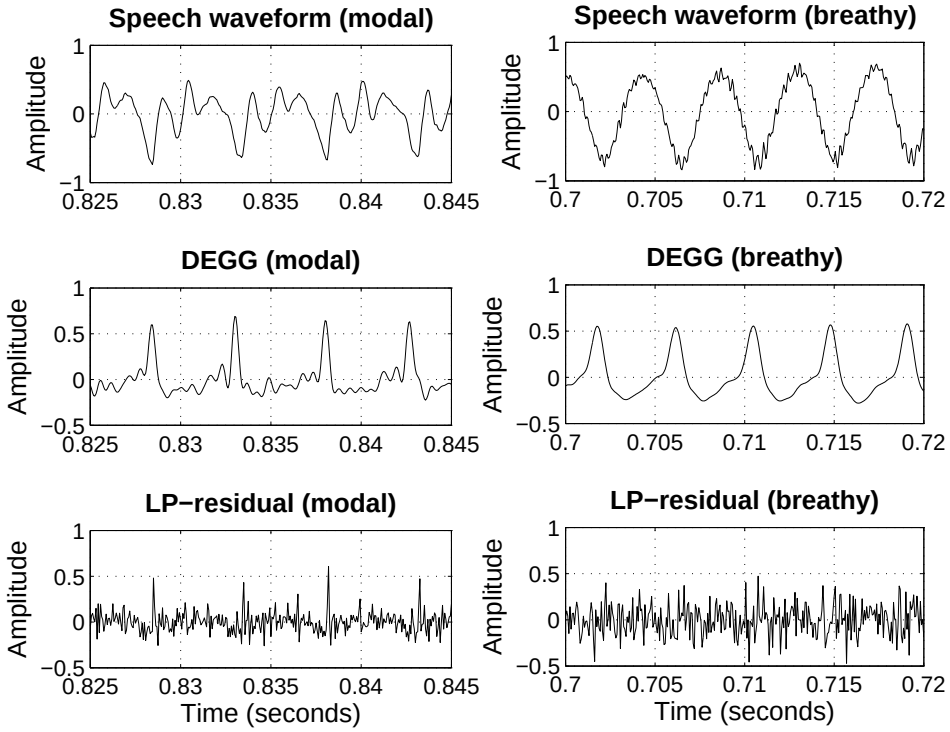


Figure 4.1: Speech waveform (top row), derivative EGG signal (middle row) and LP-residual (bottom row) for modal (left column) and breathy (right column) utterances. Speech segment is a portion of the utterance *This was easy for us* produced by a female speaker.

frequencies.

For harsh voice (right column of Figure 4.2), on the other hand, the peaks in the LP-residual are considerably less prominent, compared to both the modal and tense examples. Harsh voice is known to display irregularities in terms of amplitude and frequency of successive glottal pulses (Ishi et al., 2008a). The weaker periodicity compared to modal and tense voice combined with a noisier LP-residual (possibly contributed to by the increased presence of turbulent air produced in harsh voice, Esling and Harris, 2003) could explain the lack of prominent residual peaks observed in Figure 4.2.

Creaky voice is characterised auditorily by a sensation of repeating individual impulses (Ishi et al., 2008b) and has been studied in connection to turn-taking (Ogden, 2001) and hesitations (Carlson et al., 2006) in spoken conversation. To the best of the authors knowledge, GCI detection has not been quantitatively evaluated on creaky speech segments, although some qualitative observations were included in Degottex et al. (2009). Creaky voice produces dramatically different characteristics compared to those of modal voice (see Figure 4.3).

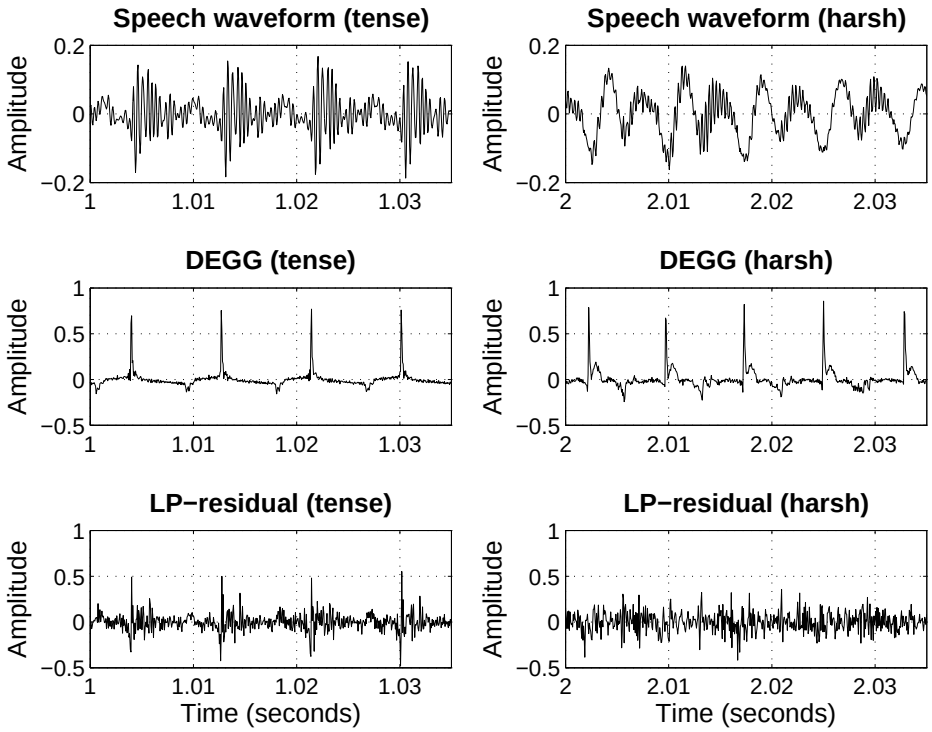


Figure 4.2: Speech waveform (top row), derivative EGG signal (middle row) and LP-residual (bottom row) for tense (left column) and harsh (right column) utterances. Speech segment is a portion of the utterance *Where were you while we were away?* produced by a male speaker.

The characteristically long pulses can be observed in Figure 4.3 as well as the presence of secondary excitations which are apparent in the DEGG signal (middle panel). Secondary peaks can also be observed in the LP-residual (bottom panel), but these peaks do not appear to correspond to the DEGG secondary peaks. Instead they are likely due to the discontinuity at glottal opening, as they are aligned to the small negative peaks in the DEGG signal. This discontinuity can be particularly abrupt in creaky voice and causes the positive peaks in the LP-residual.

Falsetto is a phonation type typically associated with high pitch and is exploited as an expressive tool in spoken conversation, frequently associated with fear (Schroeder, 2001) and femininity (Podesva, 2007). Aside from the short glottal pulses typically observed in falsetto (van den Berg, 1968), it is also reported to display a steeper spectral slope (Monsen and Engebretson, 1977). A segment of a modal utterance compared to a falsetto one by the same female speaker is shown in Figure 4.4. The difference in fundamental frequency is clearly apparent when observing the number of pulses in the two speech segments (both segments are 20 ms). The modal utterance has a median f_0 of just under 200 Hz, while

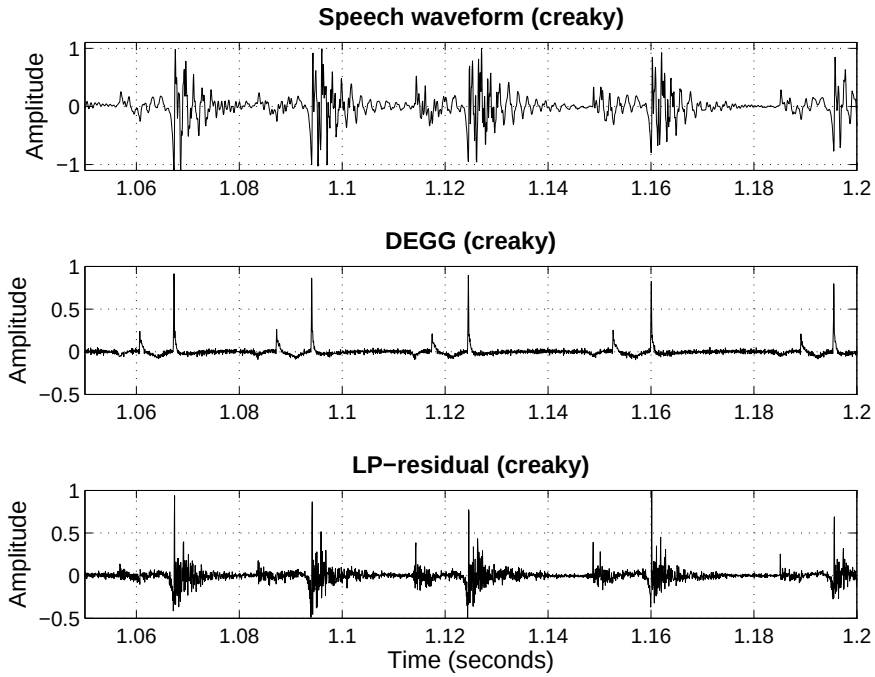


Figure 4.3: Speech waveform (top panel), derivative EGG signal (middle panel) and LP-residual (bottom panel) of a portion of the utterance *She is thinner than I am* produced by a male speaker in creaky voice.

the falsetto utterance has a median f_0 of over 400 Hz. Again for the modal utterance clear peaks occur in the LP-residual which correspond to the negative zero-crossing in the DEGG signal. For falsetto, the LP-residual is considerably noisier. For higher f_0 levels LPC analysis can become biased towards the harmonics and the subsequent inverse filtering may remove excitation components of the signal. This can contribute to a LP-residual with less prominent peaks.

From the above description it is clear that certain phonation types can display highly different characteristics from those of modal voice. As a result the strategies employed in GCI detection algorithms may not be suited to the analysis of non-modal phonation.

4.3 GCI algorithms

The automatic detection of GCIs has received a considerable amount of attention and many innovative approaches using a wide range of signal processing techniques have been proposed. For instance, some approaches rely on the Hilbert Envelope derived from the LP-residual in order to remove the finer structure in the signal and make peaks corresponding to GCIs possible to detect (Ananthapadmanabha and Yegnanarayana, 1979;

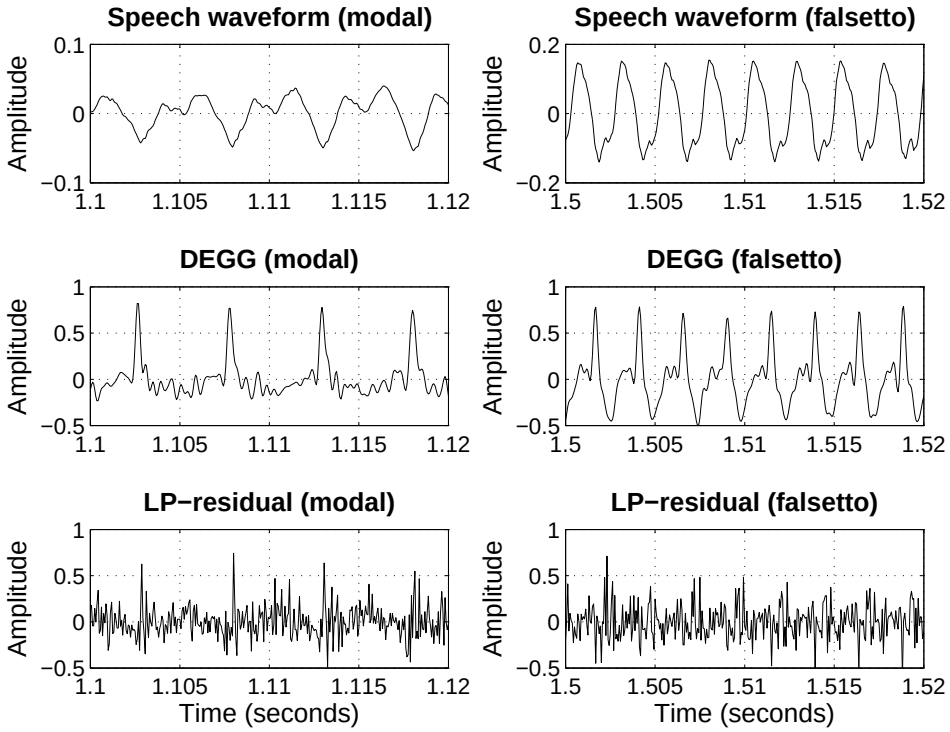


Figure 4.4: Speech waveform (top row), derivative EGG signal (middle row) and LP-residual (bottom row) for modal (left column) and falsetto (right column) utterances. Speech segment is a portion of the utterance *Where were you while we were away?* produced by a female speaker.

Cheng and O'Shaughnessy, 1989; Rao et al., 2007). Some researchers have utilised the wavelet transform, which is known to be suitable for finding singularities in signals, in order to detect GCIs (see, for example, Kadambe and Bourdreaux-Batels, 1992; Tuan and d'Alessandro, 199; Sturmel et al., 2009; d'Alessandro and Sturmel, 2011). The approach proposed in Schnell (2007) is to use a weighted, non-linear prediction method in order to derive a signal from which maxima, corresponding to GCIs, can be measured. Methods are also outlined in Moulines and Di Francesco (1990) for detecting GCIs based on sudden changes in spectral characteristics within a glottal period.

A review is presented in this section of a range of GCI detection algorithms which are evaluated in the current study. State-of-the-art algorithms were selected that both represent a wide range of approaches and whose code was available as original implementations. The code was either available online or was provided by the authors. All algorithms were used with the standard default settings and no voicing decision was carried out internally by any of the algorithms. Instead the Summation of Residual Harmonics (SRH) method (Drugman et al., 2011) was used to detect unvoiced regions. The SRH method extracts an

f_0 contour following a summation calculation carried out on the harmonics of the residual signal. The voicing decision is made by applying a threshold to the SRH value, and has been shown to be robust. All GCIs detected by the various algorithms found in the unvoiced regions were removed.

4.3.1 DYPSA

The Dynamic Programming Projected Phase-Slope Algorithm was originally presented in Kounoudes et al. (2002), with a more thorough description available in Naylor et al. (2007). It is a frequently used comparison algorithm and the original implementation is available in the VOICEBOX Matlab toolbox. The algorithm uses the phase-slope function, where positive-going zero-crossings are selected as GCIs. The phase slope function is calculated from the phase-slope, $\tau(\omega)$, which is derived from the group-delay function $-\phi(\omega)$ of the LP-residual signal (calculated from pre-emphasised speech). Zero-crossings are used as GCI candidates, and a further operation of phase slope projection is used in order to retrieve missed GCIs, which are then inserted at likely time locations.

As the window size used can greatly influence the zero-crossings found an N-best dynamic programming algorithm (Schwarz and Chow, 1990) is used to select the most suitable sequence of GCI estimates, given a specified cost function. The cost function consists of five elements which consider: inter-pulse similarity, pitch deviation, costs derived from the projected phase-slope, normalised energy values and deviations from an ideal phase-slope function. Each of these elements are weighted with constant values which are justified in Naylor et al. (2007). The N-best dynamic programming algorithm is used to find an optimal subset of the total set of GCI candidates which produce the cheapest path and thus removing false alarms.

4.3.2 ESPS

A method for detecting GCIs was described in Talkin (1989) which uses similar approaches to those used for extracting f_0 contours in the widely used *get_f0* algorithm (Talkin, 1995). First, information from *get_f0* is used to highlight unvoiced areas which are not to be analysed. The method uses dynamic programming to select the best GCI candidates which are measured as maxima in the short-term energy normalised LP-residual. Target costs consider peak amplitudes in the LP-residual and transition costs are based on periodicity and inter-pulse similarity.

Although not widely used in many recent GCI evaluation studies, this method was found to have comparable performance to state-of-the-art methods in a recent study by

Cabral et al. (2011a) and is, hence, included in the evaluation here. The algorithm is available in the ESPS/*wave+* software package.

4.3.3 YAGA

The Yet Another GCI Algorithm (YAGA, Thomas et al., 2012) combines several different methods used in other GCI algorithms, including: wavelet analysis, the group delay function and N-best dynamic programming. The method can also be used for estimating glottal opening instants (GOIs).

GCI candidates are detected in YAGA by first estimating the voice source signal by using Iterative Adaptive Inverse Filtering (IAIF, Alku, 1992). The multi-scale product of the stationary wavelet transform (SWT) is used to highlight discontinuities in the voice source signal, by taking information across the wavelet scales. These discontinuities are detected using the group-delay function, and GCI candidates are measured as negative-going zero-crossings.

False alarms are then removed using a similar N-best dynamic programming approach as is used in DYPSA (Naylor et al., 2007). YAGA uses similar cost elements to those used in DYPSA, with modification of the inter-pulse similarity cost and a further cost for distinguishing GCIs and GOIs.

4.3.4 ZFF

A GCI detection method was presented in Murty and Yegnanarayana (2008) which is based on the use of a Zero-Frequency Filter (ZFF). The approach is based on the observation that, similar to an impulse excitation, the discontinuity which occurs at GCIs is reflected across all frequencies, including 0 Hz. This information can be retrieved through the use of a 0-Hz resonator, which has the advantage of isolating frequencies well below possible vocal tract resonance.

The method has essentially four steps. The speech signal, $s(n)$, is differenced in order to remove any time varying low frequency bias, with the output being $x(n)$:

$$x(n) = s(n) - s(n-1) \tag{4.1}$$

$x(n)$ is then put through a resonator centred on 0 Hz (the filtering is done twice to ensure a sharper roll-off outside the frequency range of interest). The following difference equation is used:

$$y(n) = -\sum_{k=1}^{2} a_k y(n-k) + x(n) \tag{4.2}$$

where $a_1 = -2$ and $a_2 = 1$. The output signal, $y(n)$, can display a trend where it drifts away from the zero-line. To address this the mean of $y(n)$ is subtracting from the signal mean every 10 ms, which has the effect of forcing the signal to oscillate about the zero-line:

$$y_{rem}(n) = y(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y(n+m) \tag{4.3}$$

where $2N + 1$ is 10 ms in samples. Finally, the time points of the positive zero-crossings are then used to locate the GCIs as is shown in Figure 4.5.

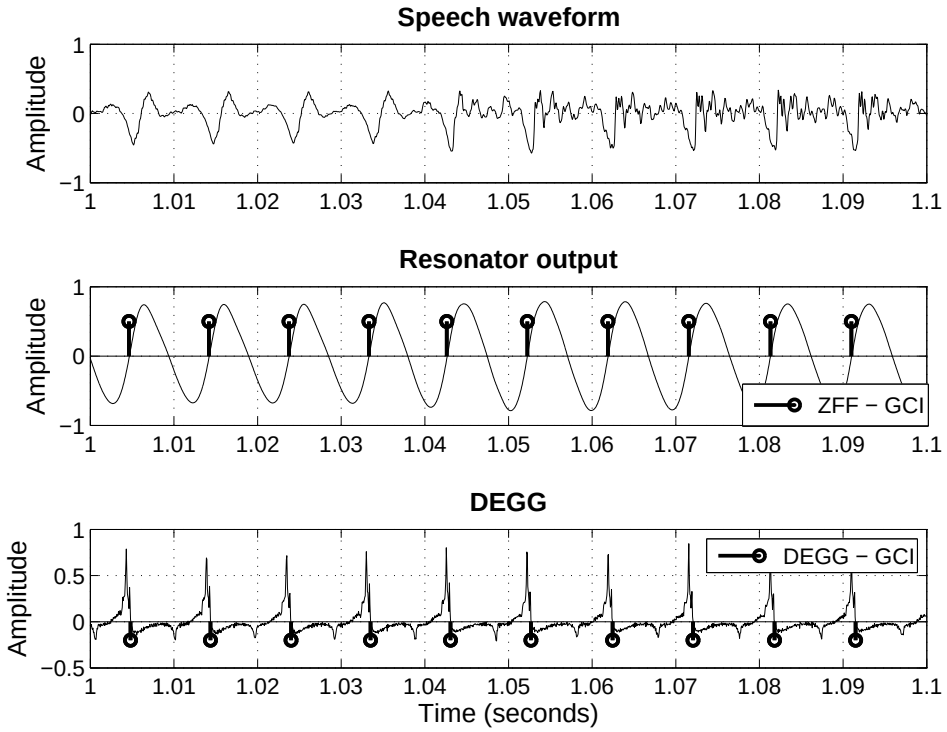


Figure 4.5: Speech waveform (top panel) of a segment of the sentence spoken by a male speaker, with the trend removed and 0-Hz filtered signal and estimated GCIs (middle panel) and the DEGG with reference GCIs (bottom panel).

4.3.5 SEDREAMS

The Speech Event Detection using the Residual Excitation and A Mean-based Signal (SEDREAMS) method was recently introduced (initially in Drugman and Dutoit, 2009,

and more comprehensively in Drugman et al., 2012c). The method uses a mean-based signal, in a similar way to the ZFF method but calculated directly from the speech signal, rather than from the output of a 0-Hz resonator. The mean-based signal, $y(n)$, is calculated from the speech signal $s(n)$ using:

$$y(n) = \frac{1}{2N+1} \sum_{m=-N}^{N} w(m)s(n+m) \tag{4.4}$$

where $w(m)$ is a Blackman window function of length $2N + 1$. This is used to determine a region where peaks corresponding GCIs in the LP-residual can be measured. It was shown in Drugman and Dutoit (2009) that window length affects the performance of SEDREAMS. Too short a window causes unwanted extra oscillations which introduces false alarms and too long a window results in misses because of oversmoothing. As a result $1.75 \cdot \frac{fs}{f_{0,mean}}$ is the chosen window size, where fs is the sampling frequency and $f_{0,mean}$ is the mean fundamental frequency, which is extracted in this study using the Summation of Residual Harmonics (SRH) method (Drugman et al., 2011).

After calculating $y(n)$, intervals are defined as the region from each minimum to a length 0.35 times the current pulse length. Maxima in the LP-residual are then measured in these regions in order to localise the GCIs.

4.4 Proposed method (SE-VQ)

The new method, referred to as SE-VQ, is described in this section. SE-VQ stands for the SEDREAMS algorithm (SE) modified to better handle voice qualities (-VQ) resulting from different phonation types. The structure of the proposed method is illustrated in Figure 4.6. The modifications to SEDREAMS involve applying a dynamic programming method to select the optimal path on GCI locations based on both the strength of the LP-residual peak and a further transition cost, which considers the transition from one GCI location to the next. Furthermore, a final post-processing method is included to remove the false positives that occur in creaky regions. A detailed description of the method now follows.

To determine the regions from which the GCIs are to be detected a mean-based signal is used, which is derived in the same manner as for SEDREAMS (see Section 4.3.5). To extract the GCIs from these regions, the LP-residual is exploited which is computed using autocorrelation LPC (as mentioned in Section 4.2 the order is determined by the sampling frequency, fs, according to $fs/1000+2$). Unlike in SEDREAMS where a single maximum

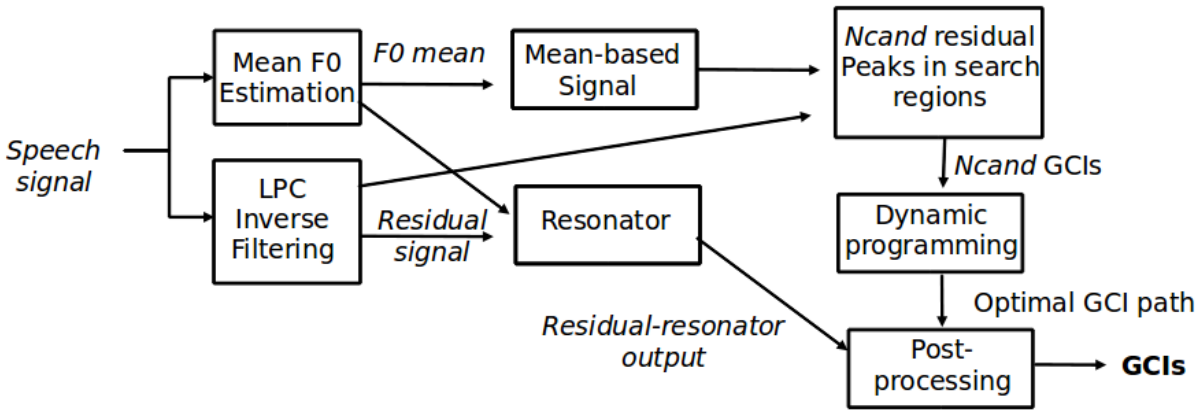


Figure 4.6: Block diagram of the SE-VQ method for detecting GCIs.

peak is chosen, several LP-residual peaks within each search region are retained. This is done in order to better handle instances where there are no prominent LP-residual peaks, e.g., in breathy or harsh voice.

For each interval i, where $1 \leq i \leq M$, of the M number of intervals detected using the mean-based signal, the N_{cand} LP-residual peaks with the strongest amplitude (in the time domain) are considered. From this a peak vector $p_{i,j}$ is produced, for $1 \leq j \leq N_{cand}$. For each value in $p_{i,j}$ a target cost $d_{i,j}$ is attributed, which is defined as:

$$d_{i,j} = \left(1 - \frac{p_{i,j}}{\max_j\{p_{i,j}\}}\right) \cdot w_p \tag{4.5}$$

which is weighted with a constant w_p. w_p is set by optimising accuracy on a database with reference GCIs (see Section 4.5.5).

The target costs are calculated for the N_{cand} peaks in each of the detected intervals, producing a structure as is shown in Figure 4.7. A transition cost, $\delta_{i,j,k}$, is here defined as:

$$\delta_{i,j,k} = (1 - |cor\{seg_{i,j}, seg_k\}|) \cdot w_s \tag{4.6}$$

where $cor\{.\}$ is the Pearson correlation coefficient of two segments of speech. $seg_{i,j}$ refers to the segment of speech centred on the j-th candidate GCI at frame i, with seg_k being the segment of speech centred around the previously chosen GCI and w_s is a constant weight (w_p, w_s and N_{cand} are jointly optimised for accuracy on a database with reference GCIs, see Section 4.5.5). This transition cost is based on the assumption that the vocal tract is relatively slowly varying within a reasonably short space of time (around 20 ms), meaning that adjacent speech segments centred on a consistently chosen point (e.g., a GCI) will

show a high degree of similarity. A sudden change in the estimated GCI location, on the other hand, will mean that adjacent speech segments centred on estimated GCIs will show less similarity. As a result, higher levels of dissimilarity are more heavily penalised.

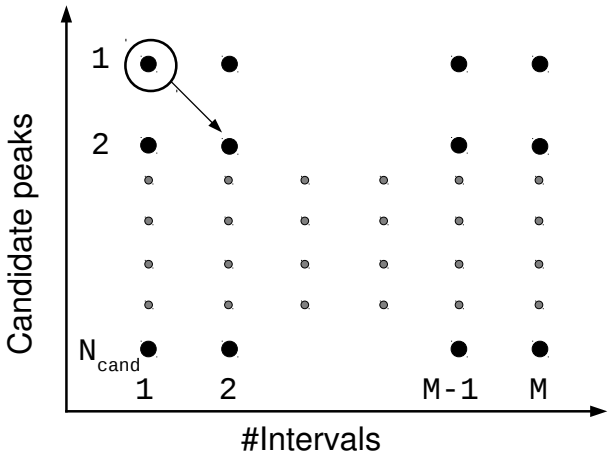


Figure 4.7: Illustration of the dynamic programming approach with N_{cand} peaks for intervals $1 \leq i \leq M$. Transition to the second optimal peak is highlighted with an arrow.

An objective function is, hence, defined incorporating the above target and transition costs. For a given frame i:

$$D_{i,j} = d_{i,j} + \min_{k \in N_{cand}} \{D_{i-1,k} + \delta_{i,j,k}\}, \quad 1 \leq j \leq N_{cand} \tag{4.7}$$

which is initialised with:

$$D_{o,j} = 0, \quad 1 \leq j \leq N_{cand} \tag{4.8}$$

Moving to the next frame the index of the optimal value is stored in a vector $q(i)$ which is used to define the optimal path of GCIs.

For many speech signals this process is typically sufficient in order to obtain suitable GCI detection. However, for creaky voice, standard GCI methods tend to display a considerable number of false positives. This is illustrated in the top panel of Figure 4.8 where the SEDREAMS method clearly produces sensible GCI detection in the non-creaky region (i.e. the beginning of the speech segment, up to 0.6 seconds) but produces a high number of false positives in the creaky region (i.e. the end of the utterance).

The final post-processing step is designed to remove these false positives, while at the same time not removing any true positive GCI estimates. To do this an analysis signal is used which is obtained by passing the LP-residual signal through a resonator centred on

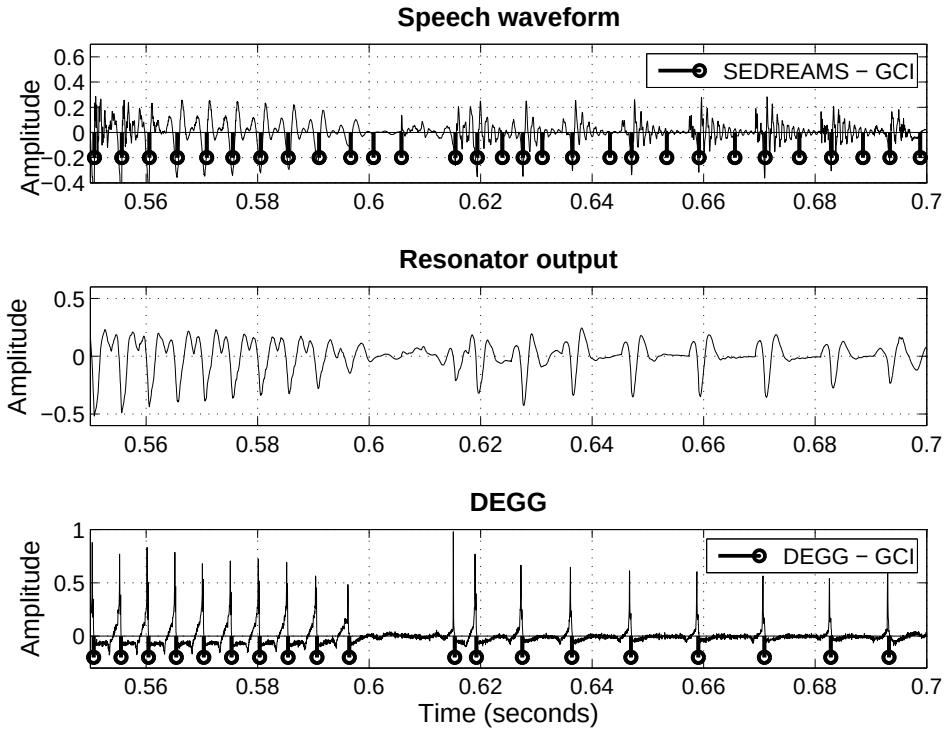


Figure 4.8: Speech waveform from an utterance produced by a male speaker with creaky voice starting from around 0.61 s and GCIs as estimated by SEDREAMS superimposed (top panel), the resonator output (middle panel) and the DEGG signal with reference GCIs (bottom panel).

$f_{0,mean}$ (see Figure 4.6). The resonator is characterised by two complex conjugate poles, the difference equation of which can be written as:

$$y(n) = A \cdot r_{LP}(n-1) + B \cdot y(n-1) + C \cdot y(n-2) \tag{4.9}$$

where

$$A = 1 - B - C \tag{4.10}$$

$$B = 2 \cdot e^{-\pi B_w T_s} \cdot cos(2\pi f_{0,mean} T_s) \tag{4.11}$$

$$C = -e^{-2\pi B_w T_s} \tag{4.12}$$

where T_s is the sampling period ($\frac{1}{F_s}$), B_w is the resonator bandwidth and r_{LP} is the Linear Prediction residual. The filtering is carried out in a forwards-backwards manner to maintain the original phase spectrum of the input signal. The resonator bandwidth, B_w, is set to 155 Hz in order to have a sufficiently pronounced resonating character to

help emphasise the most prominent residual peaks. An example of the resonator output is shown in the middle panel of Figure 4.8.

It can be observed that the resonator output displays strong negative peaks near likely GCI locations both in creaky and non-creaky regions of the speech segment. Interestingly, at detected GCI locations in the creaky region which are likely to be false positives strong negative peaks do not occur. It is this specific characteristic that is exploited in the post-processing algorithm.

The method works by iterating through the estimated GCI set and at each step it considers whether the current GCI should be removed or not. For the current GCI the corresponding negative peak in the resonator output is measured. The post-processing step is designed to remove GCIs where there are no corresponding negative peaks in the resonator output. It is illustrated with the following pseudocode:

Algorithm 4.4.1: POST-PROCESSING TO REMOVE 'FALSE' GCIS IN CREAKY REGIONS()

if $\left(\frac{r_{LP}(i-1)+r_{LP}(i+1)}{2}\right) \cdot w_{pp} > r_{LP}(i)$
then REMOVE $GCI(i)$
else RETAIN $GCI(i)$

where i is the index of the current estimated GCI, $r_{LP}(i)$ is the amplitude of the strongest negative resonator peak in the 2 ms vicinity of the current estimated GCI and w_{pp} is a constant weight (see further Section 4.5.5). The method then continues to the next estimated GCI. The remaining estimated GCIs are then considered to be the algorithms outputted set.

4.5 Evaluation

4.5.1 Speech data

Standard database

An initial test aimed at evaluating GCI detection performance of the various algorithms on standard read speech databases (as is done in the majority of GCI studies). For this speech data from a Canadian male speaker (**JMK**) and an American female speaker (**SLT**) from the ARCTIC database (Kominek and Black, 2004) was used. The JMK database contains 1114 sentences and SLT contains 1132 sentences. The sentences were designed to be phonetically balanced and simultaneous electroglottographic (EGG) signals are available

for both databases. Both audio and EGG signals, originally sampled at 32 kHz, were downsampled them to 16 kHz using the SOX toolkit.

Voice quality database

The next step aimed at evaluating GCI detection on voice qualities stemming from a range of phonation types. However, no public databases are available which contain a large volume of speech produced in a range of voice qualities and that also have simultaneous EGG recordings. Therefore, new recordings were made including audio and EGG signals. This consisted of speech produced by 6 speakers (3 male and 3 female, all experienced in speech research) recorded in a semi-anechoic chamber. This set of recordings is referred to as **read-VQ**. Audio was captured using high quality recording equipment: a B & K 4191 free-field microphone and a B & K 7749 pre-amplifier. The microphone was placed at a distance of approximately 30 cm from the speaker and participants were asked to keep this distance as constant as possible throughout the recording session. A standard laryngograph device was used to capture EGG signals, which was worn around the neck of the participants with electrodes placed on either side of the larynx. The signals were digitised at 44.1 kHz (using a LYNX-two sound card) and were subsequently downsampled to 16 kHz using the SOX toolkit.

Participants were asked to read 17 sentences in six different phonation types (making a total of 612 sentences) [3]. The sentences were chosen from the *phonetically compact* sentences in the TIMIT corpus, four of which contained all-voiced sounds. These sentences were chosen in order to obtain a wide phonetic coverage, and as it is likely that it can be very difficult for speakers to maintain a constant type of phonation over a long utterance, the short, *phonetically compact* sentences of the TIMIT corpus were selected. One of the male speakers produced 10 extra *phonetically compact* sentences, in each of the phonation types. These were used in the weight training process described in Section 4.5.5.

The set of phonation types was: breathy, modal, tense, harsh, creaky and falsetto. During the recordings, however, many of the participants found it very difficult to produce creaky voice consistently across an entire utterance. Therefore the creaky utterances were handled separately from the rest (see Section 4.5.2). Excluding creaky sentences, the voice quality dataset totalled 510 sentences.

Participants were given prototype voice quality examples, produced by John Laver [4] and the present author, and were asked to practise producing them before coming to

[3] Samples from the recordings are available at: `http://www.tcd.ie/slscs/postgraduate/phd-masters-research/student-pages/johnkane.php`

[4] These examples come as part of Laver (1980)

the recording session. For the recordings participants were asked to produce the strong versions of each phonation type and to maintain it throughout the utterance. During the recording session participants were asked to repeat the sentence when it was deemed necessary.

Although the recording process precluded the elicitation of voice quality variation as occurring in conversational speech, this process was necessary in order to obtain a large amount of data for each phonation type with good phonetic coverage. Nevertheless, it is believed that the phonation types recorded will present a realistic range of acoustic characteristics and, hence, should be suitable to test the performance of the different algorithms.

4.5.2 Creaky utterances

As participants found it particularly difficult to produce creaky voice for entire sentences a selection of creaky utterances was made. First participants were selected who were deemed to produce good renditions of creaky voice. Three of the participants, 2 male and 1 female, were included. Manual annotations were then carried out of these sentences in order to select only the regions where creaky voice was truly produced. To do this annotations were carried out by the present author who followed the same procedure as is used in Ishi et al. (2008b). The decision on creaky regions was ultimately based on an auditory judgment. The auditory criterion used was "a rough quality with the sensation of additional impulses" (Ishi et al., 2008b), but spectrograms and pitch contours were also used to help guide the annotation. Following this, annotations were given to a colleague, also experienced in voice quality research. Any regions where there was disagreement or uncertainty over the annotation were removed from the analysis. This procedure resulted in a set of 41 sentences for which creaky regions were annotated.

Also included were creaky voice data which were not directly elicited. However, very few speech databases are available which both contain creaky voice and have simultaneous EGG recordings. Note also, that the collection of such data is rather difficult, but fortunately, an American English speaker (labelled **BDL**) in the ARCTIC speech databases (Kominek and Black, 2004) happens to regularly produce creak in parts of his read sentences. 100 sentences were selected from this database which were deemed to contain creak within part of the utterance. The same annotation procedure as above was carried out on these 100 sentences.

A summary of the speech data used in this study is shown in Table 4.1.

Table 4.1: Summary of speech data used in this study.

Purpose	Database	Speakers	No. of sentences
Evaluation	JMK	Male	1114
	SLT	Female	1132
	Read-VQ	3 Male, 3 Female	510
	Read-VQ (creaky)	2 Male, 1 Female	41
	BDL	Male	100
Development	Read-VQ	Male	60

4.5.3 Perceptual evaluation

Despite attempting to ensure that each phonation type and resulting voice quality was satisfactorily produced during the recording of the **read-VQ** set, it was not believed that this was sufficient to be certain that the recorded speech data were as was instructed. As a further screening of the speech data a perceptual evaluation of the data was carried out with two participants who were experienced with using the Laver labelling scheme (note that creaky utterances and the BDL database were not included in this procedure). This was done using a web-based application where participants were randomly presented with the recorded utterances and had to choose a label, from the list: breathy, modal, tense, harsh, falsetto or other. They were also asked to state whether they believed the chosen label was produced: for *some*, for *most* or *throughout* the utterance, and whether they were: *very* confident, *quite* confident, or *not* confident that the label they had chosen was suitable.

For the current study only utterances where both listeners selected the same label, perceived the voice quality to have been sustained *throughout* the utterance, and were *very* confident with the label they had chosen were included. The purpose of this was to obtain more reliable labelling. Figure 4.9 shows the percentage of utterances, for each label, which were included in the current study. This demonstrates the importance of carrying out the perceptual screening process, particularly as less than 60 % of harsh and tense utterances were retained for analysis.

4.5.4 Reference GCIs and evaluation metrics

In order to quantitatively evaluate GCI detection performance, *reference* GCIs were extracted from the derivative of the EGG signal (DEGG). The DEGG signal displays very clear strong peaks across the range of phonation types considered here (see Figures 4.1-

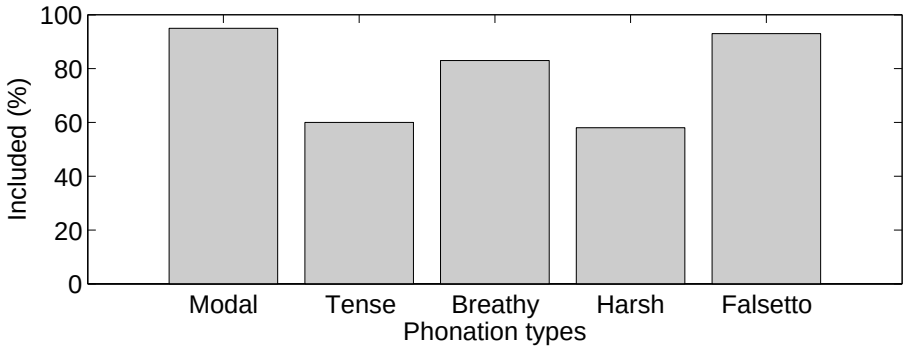


Figure 4.9: Percentage of the total recorded utterances from the **read-VQ** database that was included in the analysis, for each phonation type.

4.4). Although some algorithms have been proposed for detecting these peaks (e.g., the *pitchmarks* method from Edinburgh Speech Tools, or the SIGMA algorithm, Thomas and Naylor, 2009), instead peaks were detected within a percentage threshold of the maximum DEGG peak for a given utterance. A similar approach is used in the early stages of the DECOM algorithm (Henrich et al., 2004) and has also been used in recent GCI studies (Drugman et al., 2012c). For the current study peaks were detected above 10 % of the maximum DEGG peak for a given utterance, and the location of these peaks was used as the GCI reference. For creaky utterances, however, a slightly higher threshold of 20 % was used, in order to avoid detecting secondary excitations (as were seen in Figure 4.3, middle panel). Each creaky DEGG signal was checked to ensure that no 'true' GCIs were excluded. However, as creaky utterances in the dataset produced rather prominent DEGG peaks no 'true' GCIs were found to be excluded in the annotated creaky regions. This separate criterion is believed to be justified as a 10 % criterion would detect some of the secondary DEGG peaks as GCIs and not others. Furthermore, if a GCI is to be considered 'the most significant excitation within each glottal pulse' then selection of these secondary peaks, for example those in Figure 4.3, would imply f_0 values considerably higher than what has been previously reported in the literature. It is hypothesised that these secondary peaks in the DEGG result from ventricular incursion which appears to be present in Figure 4.3, though not in Figure 4.8.

DEGG signals, however, must be aligned to the speech signals to compensate for the delay between the laryngograph device and the microphone. This alignment was done manually for each speaker, and the delay within each speaker was assumed to be constant (as speakers were asked to maintain a constant distance from the microphone).

For a given utterance this set of reference GCIs is then used to evaluate GCIs estimated by a particular algorithm, using a set of metrics described in Naylor et al. (2007). For

a given reference GCI location $g_{ref}(r)$, the so-called *larynx cycle* is defined as the region from $g_{ref}(r) - (g_{ref}(r) - g_{ref}(r-1)) \cdot 0.5$ to $g_{ref}(r) + (g_{ref}(r+1) - g_{ref}(r)) \cdot 0.5$, see Figure 4.10.

Evaluation metrics are split into two groups: one which measures *event detection* and the second which assess detection *accuracy*.

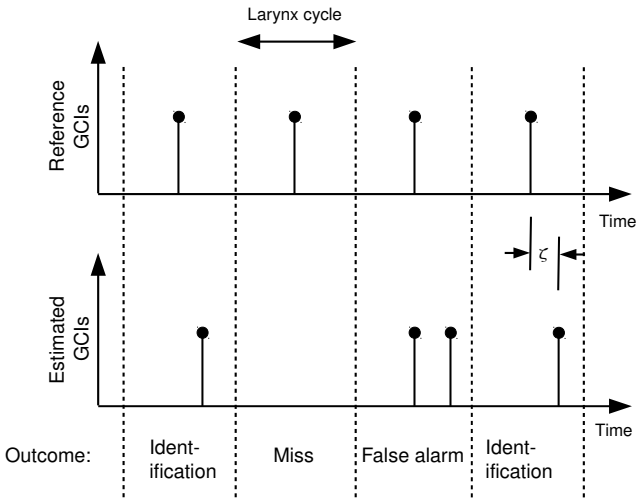


Figure 4.10: Illustration of GCI performance evaluation, given reference GCIs. Figure is based on the image shown in Naylor et al. (2007).

Identification rate (IR) is calculated as the percentage of larynx cycles where exactly one GCI is identified, i.e.

$$\text{IR} = \frac{\#\text{hits}}{\#\text{larynx cycles}} \cdot 100 \tag{4.13}$$

Miss rate (MR) is defined as the percentage of larynx cycles, in a given utterance, for which no estimated GCI is found i.e.:

$$\text{MR} = \frac{\#\text{misses}}{\#\text{larynx cycles}} \cdot 100 \tag{4.14}$$

False alarm rate (FAR) is the percentage of larynx cycles for which more than one estimated GCI is detected, i.e.:

$$\text{FAR} = \frac{\#\text{false alarms}}{\#\text{larynx cycles}} \cdot 100 \tag{4.15}$$

IR, MR and FAR all come under the *event detection* category. Note that the IR metrics is affected by misses, false alarms and correctly identified GCIs.

Identification error, ζ, is measured when exactly one estimated GCI is detected within

a given larynx cycle. ζ is the distance from the reference GCI to the estimated GCI. Identification accuracy (IDA) is measured as the standard deviation of ζ. Low IDA values imply better GCI estimation accuracy. The IDA metric was considered within the *accuracy* category.

4.5.5 Weight setting for SE-VQ

For the proposed method, SE-VQ, several constants need to be set for the experimental part of this study. These weights were determined following analysis of the set of 'Extra' sentences produced by a single male speaker as described in Section 4.5.1 (note these utterances were not included in the evaluation).

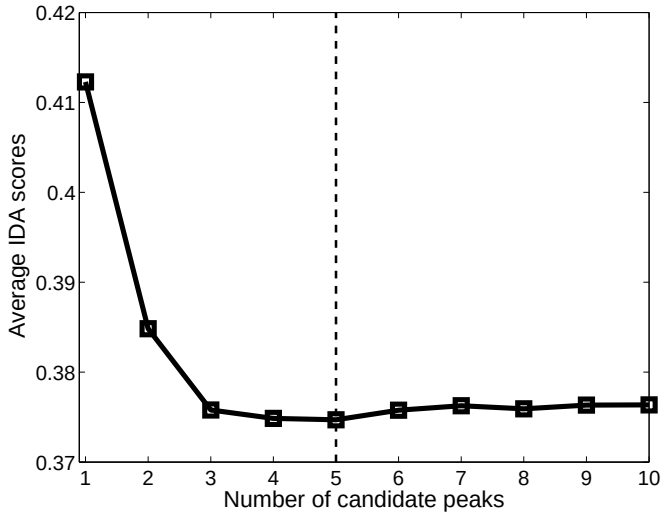


Figure 4.11: Illustration of the effect of varying the N_{cand} parameter (i.e. the number of GCI candidates considered for each detected interval) on the identification accuracy (IDA) metric. The 'Extra' sentences are the data used here, and for each datapoint the IDA value obtained for each phonation type is averaged. w_p and w_s are here set to 0.3 and 1, respectively.

For setting the weights used in the dynamic programming (i.e. w_p, w_s and N_{cand}) part, an exhaustive search was carried out by varying the three weights over a range of values and recording the identification accuracy (IDA) score for each configuration. [5] w_p and w_s, were both varied in the range [0, 1] in steps of 0.1. N_{cand} was varied in the range [1, 10], with a value of 1 essentially being the standard SEDREAMS method (excluding the post-processing procedure). The minimum IDA score was recorded for the values 0.3, 1 and 5 being set for w_p, w_s and N_{cand} respectively. The effect of varying N_{cand}, while

[5]Note that a similar weight setting approach is used in DYPSA, YAGA and ESPS.

using the optimal settings for w_p, w_s, on the metric IDA is shown in Figure 4.11. The improvement clearly starts to plateau at a value of 3, with 5 being the optimal setting.

A further weight, w_{pp}, also needs to be set. w_{pp} is the parameter of the post-processing method described in Section 4.4 which is used to remove false positives in creaky regions. w_{pp} is not jointly optimised with w_p, w_s and N_{cand}. Using the three constants already set, another exhaustive search is carried out on the 'Extra' speech data. Note that these data contain creaky utterances as well as the other five phonation types. This time w_{pp} is varied in the range [0, 1] in steps of 0.1. For each step of the search an overall identification rate (IR) score is retained. The setting 0.4 obtained the highest IR score and was therefore used in the evaluation.

4.5.6 Statistical analysis

Typically in the literature, studies present evaluation metrics derived at the speaker level, however to facilitate stable statistical analysis the metrics IR and IDA were derived at the sentence level. This was done just for the voice quality dataset. For each phonation type, one-way ANOVAs were carried out by treating the evaluation metric as the dependent variable and algorithm type as the independent variable. Subsequent post-hoc analysis (i.e. pairwise comparisons) was conducted using Tukey's Honestly Significant Difference (HSD) test.

All the evaluation metrics were also derived at the speaker level and these results are also presented in the results.

4.6 Results

4.6.1 Standard Evaluation

The first set of results, from the analysis of the ARCTIC **SLT** and **JMK** databases are presented in Table 4.2. One can observe largely similar results for the proposed method, SE-VQ, compared with SEDREAMS. Identification rate (IR) was slightly better for SE-VQ in SLT, and slightly worse in JMK. Identification accuracy (IDA) for SE-VQ was better than SEDREAMS in both databases.

SE-VQ achieved the highest IR in SLT (98.91 %) with YAGA getting the highest in JMK (99.28 %). Generally IR was very high across the six algorithms. Although DYPSA showed the lowest score in both databases (97.48 % for SLT and 98.84 % for JMK), differences were rather small. The YAGA method displayed the lowest IDA in JMK

Table 4.2: Summary of evaluation results, with metrics derived at the speaker level, for the ARCTIC **SLT** and **JMK** databases for the 6 algorithms. Best scores for each metric, within each database are marked in bold.

Database	Algorithm	IR (%)	MR (%)	FAR (%)	IDA (ms)
SLT	DYPSA	97.48	1.19	1.33	0.40
	ESPS	98.90	0.58	**0.52**	0.22
	SEDREAMS	98.71	0.31	0.98	0.26
	SE-VQ	**98.91**	0.28	0.81	0.24
	YAGA	98.55	0.36	1.09	0.21
	ZFF	98.27	**0.27**	1.46	**0.16**
JMK	DYPSA	98.84	0.33	0.83	0.51
	ESPS	99.10	0.50	**0.40**	0.52
	SEDREAMS	99.21	0.21	0.58	0.46
	SE-VQ	98.91	0.67	0.42	0.44
	YAGA	**99.28**	**0.07**	0.65	**0.34**
	ZFF	99.16	**0.07**	0.77	0.48

(0.34 ms). DYPSA showed the lowest accuracy on SLT (0.40 ms), though the ESPS and ZFF method, both of which performed well on SLT, showed slightly degraded accuracy performance in JMK.

The reader may note that the event level metrics (i.e. IR, MR and FAR) provide a different performance indication that the accuracy metric (i.e. IDA) and that the two groups are somewhat independent. For instance, it can be observed that although the ZFF algorithm does not provide the best IR (indeed it provides the second worst) it does, however, provide the best IDA. This indicates that although ZFF produces a relatively high number of false alarms (due to excessive oscillations in the resonator output), when a single GCI is detected its localisation is very precise.

The results here largely corroborate those presented in Drugman et al. (2012c) where the ZFF, YAGA and SEDREAMS algorithms outperformed DYPSA in both event and accuracy metrics. In terms of the proposed algorithm, SE-VQ, these results are encouraging as they provide evidence that the post-processing component (used for removing false positives in creaky regions) does not remove 'true' GCIs in non-creaky speech. The strong performance of the YAGA algorithm for JMK may be partly explained be improved robustness at voice onset/offset. The lower performance for the female speaker, SLT, could be due to the use of the IAIF algorithm which tends to have its effectiveness reduced for

higher pitched voices.

4.6.2 Voice quality database

A summary of the evaluation metrics (derived at the speaker level) is presented for the **read-VQ** recordings in Table 4.3. Metrics were obtained by each of the algorithms, for each phonation type and averaged across the six speakers. The distributions of IDA and IR metrics (derived at the sentence level) are plotted as a function of algorithm type and for each phonation type in Figures 4.12-4.13 and 4.14-4.15, respectively. For the creaky category, however, only the three selected speakers (with annotated creaky regions) are represented in Table 4.5. In Figures 4.13 and 4.15 the creaky category also contains metrics obtained from the BDL database. Note that the results for creaky voice will be considered in Section 4.6.3.

Comparing the proposed method, SE-VQ, to SEDREAMS one can observe an improvement in mean identification accuracy (IDA) across all phonation types. Considering the IDA distributions in Figure 4.12 the largest difference was for tense voice and this was found to be significant ($p < 0.05$). In terms of event detection metrics there was little change, except for some improvement in the mean identification rate (IR) for harsh voice and falsetto. The results for each phonation type are now described separately.

For **modal** voice the ESPS (0.29 ms), SEDREAMS (0.3 ms), SE-VQ (0.26 ms) and ZFF (0.25 ms) algorithms provided the lowest mean IDA scores. ESPS and YAGA, however, displayed relatively high levels of variance. DYPSA algorithm displayed significantly worse ($p < 0.001$) IDA and IR scores than all the other algorithms. SEDREAMS (99.48 %) and SE-VQ (99.41 %) algorithms providing high IR values and lowest variance. DYPSA and ZFF produced the highest False Alarm Rate (FAR). These results follow a similar trend to that seen in Section 4.6.1.

For **tense** voice ESPS (0.18 ms), SE-VQ (0.18 ms) and ZFF (0.17 ms) gave the lowest average IDA scores. DYPSA (0.35 ms), SEDREAMS (0.24 ms) and YAGA (0.83 ms) displayed both higher mean IDA values and higher levels of variance. It is interesting to note that the mean IDA scores were reduced for tense voice compared to modal voice for all algorithms with the exception of YAGA, which displayed the same IDA scores. For IR, ESPS (99.64 %), SEDREAMS (99.74 %) and SE-VQ (99.76 %) produced the highest values. The high mean Miss Rate for DYPSA (1.01 %) and ZFF (1.64 %) was the cause of the relatively low IR for these two algorithms.

For **breathy** voice the opposite trend is observed in terms of IDA scores, compared with tense voice, with all of the algorithms (with the exception of ZFF) displaying higher

Table 4.3: Summary of evaluation results, with metrics derived at the speaker level, for the 6 algorithms, separated by phonation type and averaged over the 6 speakers for the **read-VQ** recordings. Both mean ($\bar{x}$) and standard deviation (σ) are presented for the identification rate (IR) and identification accuracy (IDA) metrics. Best mean scores within each phonation type category are marked in bold.

Phonation Type	**Algorithm**	**IR (%)**		**MR (%)**	**FAR (%)**	**IDA (ms)**	
		$\bar{x}$	σ	$\bar{x}$	$\bar{x}$	$\bar{x}$	σ
Modal	DYPSA	96.30	1.82	1.55	2.15	0.58	0.31
	ESPS	99.06	0.68	0.72	**0.22**	0.29	0.09
	SEDREAMS	**99.48**	0.35	0.10	0.42	0.30	0.06
	SE-VQ	99.41	0.36	0.19	0.40	0.26	0.06
	YAGA	99.25	0.50	0.14	0.61	0.32	0.07
	ZFF	98.93	0.68	**0.04**	1.03	**0.25**	0.02
Tense	DYPSA	98.08	0.69	1.01	0.91	0.35	0.14
	ESPS	99.64	0.28	0.18	0.18	0.18	0.05
	SEDREAMS	99.74	0.26	**0.03**	0.23	0.24	0.07
	SE-VQ	**99.76**	0.25	0.04	0.20	0.18	0.07
	YAGA	99.08	1.04	0.09	0.83	0.32	0.10
	ZFF	98.34	1.33	1.64	**0.02**	**0.17**	0.03
Breathy	DYPSA	93.24	4.15	2.53	4.23	0.72	0.22
	ESPS	97.29	3.64	1.77	0.94	0.46	0.18
	SEDREAMS	**99.01**	1.37	**0.20**	0.79	0.40	0.10
	SE-VQ	**99.01**	1.39	0.21	**0.78**	0.35	0.10
	YAGA	98.82	1.40	0.26	0.92	0.42	0.09
	ZFF	76.28	20.82	22.24	1.48	**0.24**	0.08
Harsh	DYPSA	90.27	7.59	6.70	3.03	0.74	0.38
	ESPS	97.39	2.67	**1.48**	1.13	0.56	0.19
	SEDREAMS	96.53	4.38	2.19	1.28	0.58	0.21
	SE-VQ	**97.64**	3.09	1.59	**0.77**	**0.54**	0.24
	YAGA	97.43	3.22	1.55	1.02	0.66	0.21
	ZFF	82.30	27.89	16.71	0.99	0.67	0.56
Falsetto	DYPSA	80.15	20.44	18.43	1.42	0.52	0.15
	ESPS	**97.11**	2.49	1.75	1.14	**0.34**	0.10
	SEDREAMS	93.02	8.14	5.27	1.71	0.46	0.16
	SE-VQ	94.52	8.10	4.56	**0.92**	0.41	0.15
	YAGA	87.31	19.33	11.15	1.54	0.50	0.19
	ZFF	94.83	6.08	**1.01**	4.16	0.38	0.13

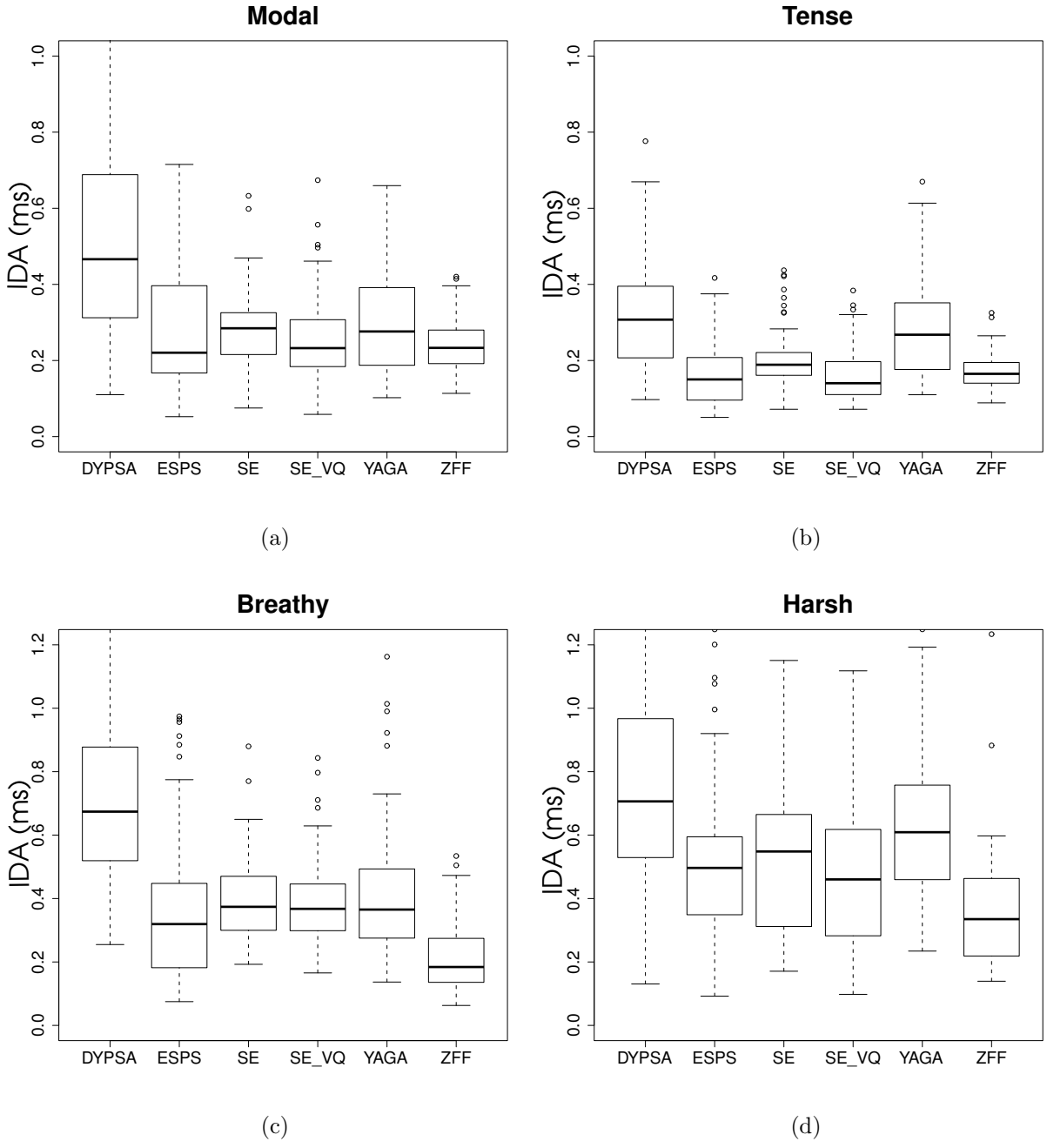
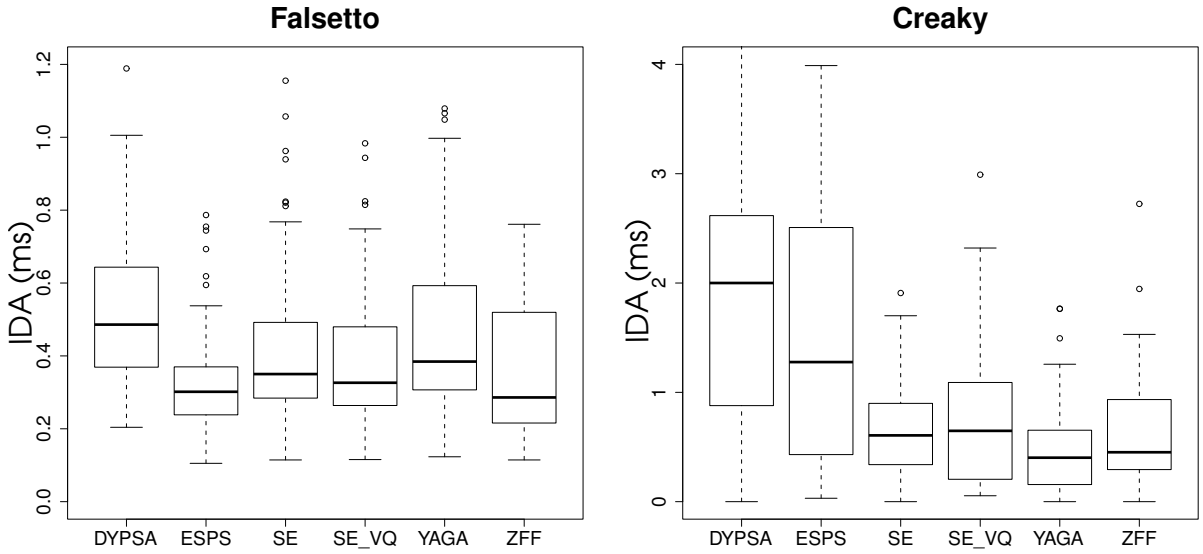


Figure 4.12: Distributions of **Identification Accuracy (IDA)**, derived at the sentence level, plotted as a function of algorithm type with separate panels for modal (a), tense (b), breathy (c) and harsh voice (d). Note that the SEDREAMS method has been abbreviated to SE for these plots.

IDA values compared to modal. ZFF (0.24 ms) and SE-VQ (0.35 ms) gave the lowest mean IDA values, with DYPSA giving significantly higher ($p < 0.001$) IDA scores than



(a) (b)

Figure 4.13: Distributions of **Identification Accuracy (IDA)**, derived at the sentence level, plotted as a function of algorithm type with separate panels for falsetto (a) and creaky voice (b). Note that the SEDREAMS method has been abbreviated to SE for these plots.

the rest. However, although ZFF had significantly lower ($p < 0.001$) IDA scores than the rest it also displayed significantly lower ($p < 0.001$) IR than the other algorithms due to a high miss rate (22.24 %). This result for ZFF was examined and it was found that the resonator output (shown in Figure 4.5), used for locating GCIs, can display a low frequency drift off the zero-line for very breathy utterances. This can at times result in the oscillating signal not producing a zero-crossing and, hence, a high miss rate. This issue may be addressed by applying further processing to the oscillating signal. At the same time, the ZFF method produced the lowest IDA (0.24 ms). Examining the 0-Hz resonator output (used in the ZFF method) for breathy utterances, the oscillations in the signal were found to closely match those in the DEGG signal (see Figure 4.5). As a result, positive-going zero-crossings in this signal produced consistently aligned GCI estimates, and hence a lower IDA. For other algorithms, e.g., SEDREAMS, the localisation of GCIs is done on the LP-residual which does not display the same regularity, and the accuracy of estimated GCIs suffers from the lack of prominent residual peaks. DYPSA also produced a low IR (93.24 %), and the ESPS method produced a larger variance compared with the SEDREAMS, SE-VQ and YAGA algorithms.

Like breathy voice, **harsh** voice also produced higher mean IDA scores, compared

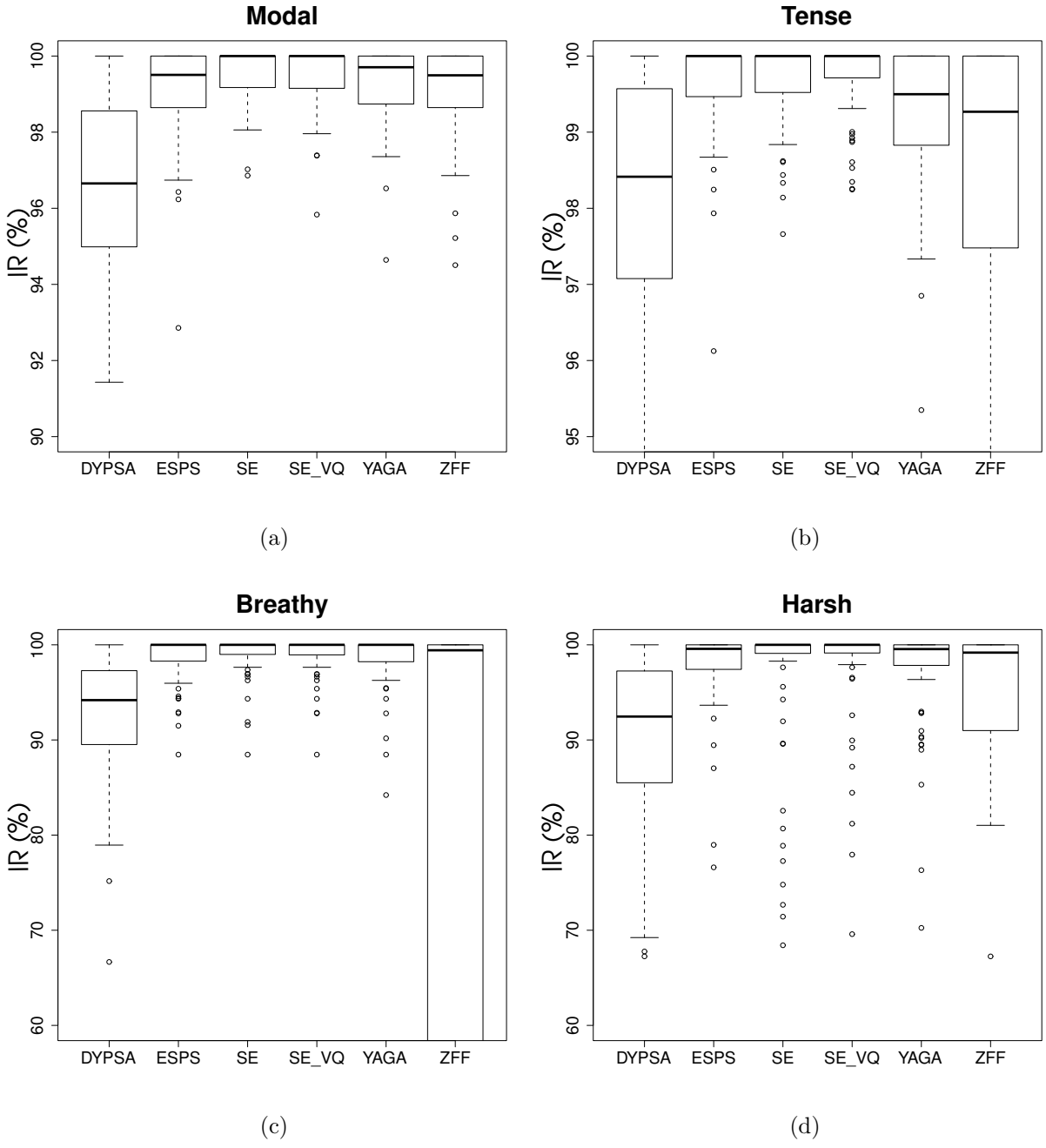
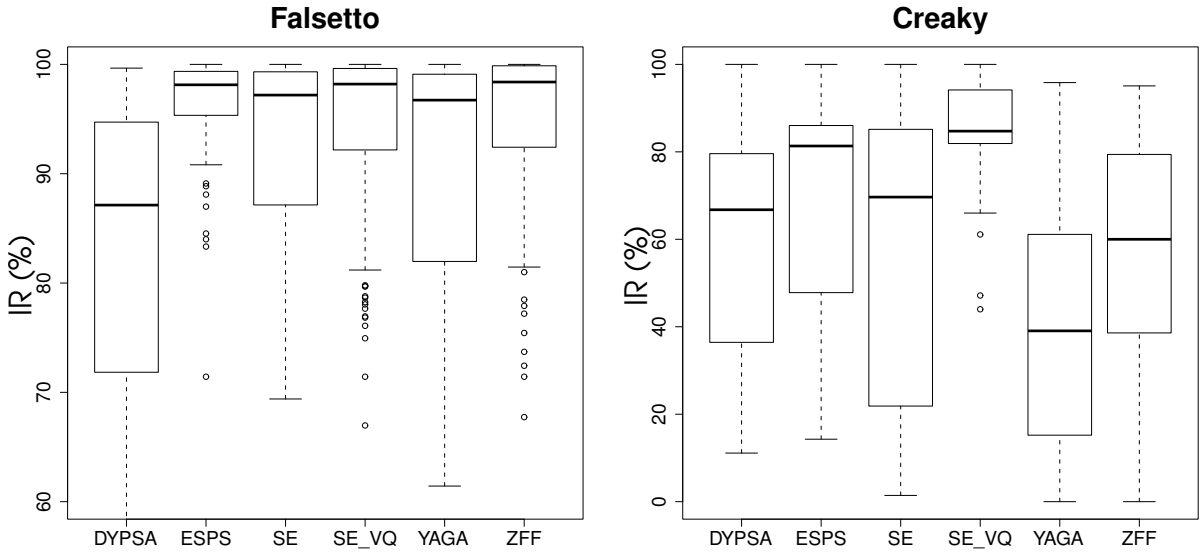


Figure 4.14: Distributions of **Identification Rate (IR)**, derived at the sentence level, plotted as a function of algorithm type with separate panels for modal (a), tense (b), breathy (c) and harsh voice (d). Note that the SEDREAMS method has been abbreviated to SE for these plots.

with modal voice, for all algorithms. The six algorithms displayed rather similar mean IDA scores, but note however in Figure 4.12d the high variance in the IDA scores for the



(a) (b)

Figure 4.15: Distributions of **Identification Rate (IR)**, derived at the sentence level, plotted as a function of algorithm type with separate panels for falsetto (a) and creaky voice (b). Note that the SEDREAMS method has been abbreviated to SE for these plots.

DYPSA and ZFF algorithms. In terms of IR, ZFF again displayed the lowest mean value (82.3 %) due to a high miss rate. As before, this was due to a drift away from the zero-axis of the resonator output. The SE-VQ algorithm provided the highest mean IR score (97.64 %) as well as the lowest mean IDA (0.54 ms).

Finally, for **falsetto** ESPS (0.34 ms), ZFF (0.38 ms) and SE-VQ (0.41 ms) displayed the lowest mean IDA values, with ESPS and SE-VQ displaying the lowest variance. Again, all algorithms showed higher mean IDA scores than those for modal voice. The highest IR scores were found for ESPS (97.11 %), ZFF (94.83 %) and SE-VQ (94.52 %). The score for DYPSA (80.15 %) and YAGA (87.31 %) were considerably lower due to a relatively high miss rate. These surprisingly low scores were investigated and it found that the f_0 ceiling is the likely cause. For instance, in both DYPSA and YAGA the maximum f_0 value is set to 500 Hz. However, in some of the falsetto utterances in the read-VQ database f_0 exceeds 500 Hz and in these cases there were frequently missed GCIs.

4.6.3 Creaky database

The results for the ARCTIC-BDL database are presented in Table 4.4 and are separated into *All* and *Creaky* speech regions. Included in Table 4.5 are the results for the annotated

Table 4.4: Summary of evaluation results for the ARCTIC **BDL** database for the 6 algorithms. Results are separated into two categories, one considering the full speech utterances and the other considering performance just in creaky voice regions. Best scores for each metric, within each category are marked in bold.

Regions	Algorithm	IR (%)	MR (%)	FAR (%)	IDA (ms)
All	DYPSA	95.83	1.73	2.44	0.56
	ESPS	98.30	1.07	0.63	0.31
	YAGA	97.80	0.34	1.86	0.30
	ZFF	86.80	**0.18**	13.02	**0.28**
	SEDREAMS	97.65	0.59	1.76	0.33
	SE-VQ	**98.86**	0.59	**0.55**	0.32
Creaky	DYPSA	82.90	2.98	14.12	1.26
	ESPS	90.16	6.76	3.08	0.66
	YAGA	82.01	0.10	17.89	0.37
	ZFF	23.76	**0.00**	76.24	0.37
	SEDREAMS	71.17	0.23	28.60	0.35
	SE-VQ	**97.32**	0.59	**2.09**	**0.28**

creaky regions in the three speakers selected from the read-VQ database (presented results are averaged across the three speakers). The distributions of IDA and IR scores for all creaky regions (i.e. both from the BDL database and the three selected speakers from read-VQ) are shown in Figures 4.13b and 4.15b, respectively.

Compared to SE-VQ, for the SEDREAMS method it was found that it produced both a lower IDA and a higher IR score for the whole BDL database, for the creaky annotated regions in BDL. Furthermore, the IR scores for SE-VQ were significantly higher than SEDREAMS ($p < 0.001$) for the read-VQ dataset. The improvement in IR is due to a considerably reduced false alarm rate, with only a slightly increased miss rate.

For the creaky regions in the BDL database the SE-VQ algorithm produced a high IR (97.32 %) compared to the five other algorithms. For ESPS, this is due to a higher miss rate. But for all other algorithms this is due to a considerably higher false alarm rate. It is then likely that this improvement on the creaky regions contributes to a general improvement in IR on the whole BDL database. SE-VQ also showed the lowest IDA for the creaky regions in BDL, whereas all the algorithms (with the exception of DYPSA) displayed a low IDA for the full BDL database.

For the creaky annotated regions in the read-VQ database SE-VQ again produced the highest IR (87.87 %). The IR for SEDREAMS was significantly high than for DYPSA

Table 4.5: Summary of evaluation results, with metrics derived at the speaker level, for the three speakers selected from the **read-VQ** database. Both mean ($\bar{x}$) and standard deviation (σ) are presented for the identification rate (IR) and identification accuracy (IDA) metrics. Best mean scores for each metric, within each category are marked in bold.

Regions	**Algorithm**	**IR (%)**		**MR (%)**	**FAR (%)**	**IDA (ms)**	
		$\bar{x}$	σ	$\bar{x}$	$\bar{x}$	$\bar{x}$	σ
read-VQ	DYPSA	63.89	16.27	5.62	30.49	1.87	0.89
	ESPS	72.94	15.56	7.77	19.29	1.57	1.08
	SEDREAMS	61.29	28.06	3.29	35.42	0.89	0.33
	SE-VQ	**87.87**	2.97	6.84	**5.29**	0.70	0.07
	YAGA	45.75	23.31	**0.55**	53.70	**0.61**	0.14
	ZFF	60.09	20.55	6.72	33.19	1.13	0.79

($p < 0.01$), YAGA ($p < 0.001$) and ZFF ($p < 0.001$), with a difference approaching significance for ESPS ($p = 0.06$). This, however, is a reduced score compared to the creaky regions in BDL. This was investigated and it was found that this is mainly due to the difficulty in estimating an $f_{0,mean}$ which is used to define the centre frequency of the resonator used in the post-processing step. Small variations in $f_{0,mean}$ have only a minor effect on the resonator output. However, large differences can have a considerable impact on the strength of the negative peaks. For the BDL utterances the creaky regions were typically in a sentence-final position. Although creaky regions can display spurious f_0 values, by calculating $f_{0,mean}$ over the whole utterance (much of which is not creaky) one can obtain a suitable setting for the resonator. However, for sentences which were completely (or almost completely) produced with creaky phonation the f_0 contour can be extremely erratic and hence the $f_{0,mean}$ value may be unsuitable for the resonator. This may explain the increase in both miss rate and false alarm rate. Nevertheless, the approach used in SE-VQ for handling creaky regions results in a higher IR than the other algorithms.

4.7 Discussion

The results in this study demonstrate that different voice qualities have a strong effect on GCI detection performance of a host of state-of-the-art algorithms. This effect is largely due to the different glottal closure characteristics of the phonation types (illustrated in Section 4.2). As expected breathy voice, harsh voice and falsetto caused overall lower GCI

detection accuracy compared to modal voice. This corroborates the initial findings presented in Cabral et al. (2011a) and some of the observations given in Tuan and d'Alessandro (199). This is believe is due to the lack of prominent peaks and other discontinuities in the analysis signals used in the various algorithms.

Apart from some qualitative observations in Degottex et al. (2009), GCI detection studies have tended not to focus on creaky voice segments. In the results, however, creaky voice clearly displayed the strongest negative effect on GCI detection performance, particularly causing a prohibitive amount of false alarms. For the proposed algorithm the post-processing component brought about a large reduction in these false alarms and produced a higher identification rate. At the same time, analysis of standard databases (containing a large number of read sentences) revealed that this post-processing step did not bring about any degradation in the performance. However, performance on utterances which involved sustained production of creaky voice was lower than for utterances with creaky voice occurring mainly in sentence-final position. Further qualitative analysis suggested that this is due to inappropriate setting of the resonator's centre frequency, following spurious f_0 values for the sustained creaky sentences. Ongoing research is looking at finding a more stable setting for this resonator to improve the detection performance on creaky utterances.

A perhaps surprising result for the proposed algorithm was that larger improvements in identification accuracy were observed for modal and tense voice compared to breathy and harsh utterances. The dynamic programming component was designed to more consistently select GCIs in cases where there was a lack of prominent peaks in the LP-residual. Further analysis was carried out and it was found that these improvements in accuracy for modal and tense voice were found to be largely at voice onset/offset. In these areas speech that is perceived as modal or tense can display characteristics typically associated with breathy voice (Ní Chasaide and Gobl, 1993). As a consequence, there will be a lack of prominent LP-residual peaks, and the dynamic programming component appeared to be suitable for addressing this. However, as the breathy and harsh voice were produced throughout the entire sentence, very few prominent LP-residual peaks were found. As a result the dynamic programming component cannot obtain a clear starting point and, hence, it is difficult to provide improved accuracy. However, in expressive and conversational speech people vary their voice quality dynamically (Gobl and Ní Chasaide, 1992). Hence, there will be regions where there are prominent LP-residual peaks, and it is hypothesised that the dynamic programming approach will be more suited to this. Further preliminary analysis has been carried out on vowels produced by changing the phonation type from modal to breathy, and some evidence to support this hypothesis has been

observed. However, this needs further, formal investigation.

Evidence from the results also demonstrate that some settings of state-of-the-art algorithms may need to be altered to handle different phonation types. Although a maximum f_0 setting of 500 Hz (as is used in DYPSA and YAGA) is unlikely to be exceeded in standard speech synthesis corpora, for expressive and conversational speech, female speakers using a falsetto voice may indeed exceed it (Podesva, 2007). Furthermore, breathy and harsh voice tended to cause the output of the 0-Hz resonator, used in ZFF, to significantly drift from the zero-line, which caused a high number of misses. Further processing of this signal could be applied to rectify this problem.

A further interesting finding was the high performance of the ESPS algorithm, compared with other state-of-the-art algorithms. The algorithm, originally presented in Talkin (1989), is considerably older than the other methods evaluated and is rarely included in GCI studies. Nevertheless, it produced consistently reliable results on the standard database as well as for many of the phonation types (particularly falsetto).

Overall, no dramatic gender effect was observed, particularly with regard to the event level metrics. There was, however, some effect on the precision of GCI localisation, with female speakers displaying generally higher (though not dramatically higher) error values. Female voices have often been described in the literature as having a breathier quality to males, which would involve less abrupt glottal closure characteristics. This may have affected the precision of the analysis.

4.8 Conclusions and future work

This chapter investigates the glottal closure characteristics of a range of phonation types in relation to the various analysis signals relevant to GCI detection. Phonation type is known to have a significant effect on a speaker's voice quality and is frequently exploited by speakers, particularly for certain expressive functions. A review is presented of the strategies employed in a range of GCI detection algorithms. A new method, SE-VQ, is outlined which involves modifications to the SEDREAMS algorithm in an attempt to deal with the potential difficulties posed by non-modal phonation types. The proposed algorithm uses a mean-based signal approach (as in SEDREAMS) for detecting intervals within which GCIs are located. A dynamic programming component was then applied to consistently select these locations, even when there were a lack of prominent peaks in the LP-residual. Finally, a post-processing component, utilising features of the LP-residual signal passed through a resonator, is applied to remove false alarms in creaky regions.

GCI detection algorithms were evaluated both on large standard databases, as well as

on speech containing a variety of phonation types. The most striking result from the study, in terms of the proposed method, was the high identification rate obtained on the creaky voice data, which was considerably higher than that of any other method. A key factor here was the post-processing method employed, which was also shown not to degrade the performance on standard databases which contained less voice quality variation.

Other phonation types, like falsetto and harsh voice, caused considerable degradation on the GCI detection compared to modal voice. This may have implications for analysis methods used to model highly expressive speech.

Future work will involve further collection of creaky voice data occurring in different contexts, with simultaneous EGG recordings. This will be used to more comprehensively evaluate GCI detection performance on this particular voice quality.

4.9 Applications

The proposed GCI detection algorithm, SE-VQ, has potential for usage in a variety of applications. In particular this method has strong potential for exploitation in parametric speech synthesis and voice modification methods that seek to produce a variety of voice types. One specific collaborative work the present author has been engaged in is modelling creaky excitations for use in a statistical parametric speech synthesis system (Drugman et al., 2012b). The new proposed method is crucial for providing a suitable modelling of the temporal dynamics of creaky voice which is necessary to produce synthetic speech with naturally sounding creaky segments. Other potential uses for the SE-VQ method include certain glottal inverse filtering methods that require precise positioning of the GCI (e.g., closed-phase inverse filtering or mixed-phase decomposition). Also, as knowledge of GCIs is a necessary prerequisite for many glottal source analysis and modelling methods (for example the one described in Chapter **??**) the usefulness of the SE-VQ for this is evident.

Relevant publications

- Kane, J., Gobl, C., (2013) Evaluation of glottal closure instant detection in a range of voice qualities, *Speech Communication*, 55(2), pp. 295-314.
- Drugman, T., Kane, J., Gobl, C., (2012) Modeling the creaky excitation for parametric speech synthesis, *Proceedings of Interspeech, Portland, Oregon, USA*.
- Cabral, J., Kane, J., Gobl, C., Carson-Berndsen, J., (2011) Evaluation of glottal epoch detection algorithms on different voice types, *Proceedings of Interspeech, Florence, Italy,* 1989-1992.

Chapter 5

Automating manual user strategies for precise glottal source analysis

Summary

A large part of the research carried out at the Phonetics and Speech Laboratory is concerned with the role of the glottal source in the prosody of spoken language, including its linguistic and expressive dimensions. Due to the lack of robustness of automatic glottal source analysis methods, the tendency has been to use labour intensive methods which require pulse-by-pulse manual optimisation. This has affected the feasibility of conducting analysis on large volumes of data. To address this, a new method is proposed for automatic glottal source parameterisation by simulating the strategies used in the manual optimisation approach. The method involves a combination of exhaustive search, dynamic programming and optimisation methods, with settings derived from analysis of previous manual glottal source analysis. A quantitative evaluation demonstrated clearly closer model parameter values to the reference values, compared with a standard time domain-based approach and a phase minimisation method. A complementary qualitative analysis illustrated broadly similar findings, in terms of glottal source dynamics in various placements of focus, when using the proposed algorithm compared with a previous study which employed the manual optimisation approach.

5.1 Introduction

A specific research focus at the Phonetics and Speech Laboratory in Trinity College Dublin is the role of the glottal source in speech. This on-going work is dedicated both to descriptive studies regarding the prosody of the voice as well as the development of more robust, automatic algorithms. Of particular interest is the role of the voice in the prosody of speech, i.e. how dynamic, temporal variation of the entire glottal source (f_0 and phonation quality), provides both the underlying linguistic prosody as well as its expressive dimension. As part of this endeavour, previous work has looked at the glottal source correlates of focus and deaccentuation (Yanushevskaya et al., 2010; Ní Chasaide et al., 2011). These studies have involved the use of labour intensive methods which require pulse-by-pulse manual fine-tuning to ensure precise glottal source characterisation. This chapter looks at methodological developments, drawing on data where utterances were elicited with different focal accentuation patterns. A new method is proposed for parameterisation of the estimated glottal source derivative that helps overcome some of the difficulties that arise with standard automatic analysis methods, by simulating the strategies used in the manual fine-tuning approach.

To derive an estimate of the differentiated glottal source signal one can first consider the speech production process (in the frequency domain) as:

$$S(f) = G(f)V(f)L(f) \tag{5.1}$$

where the spectrum of the speech output, $S(f)$, is the product of the three components $G(f)$, $V(f)$ and $L(f)$, where $G(f)$ is the spectrum of the glottal flow signal (i.e. the glottal source spectrum), $V(f)$ is the transfer function of the vocal tract, $L(f)$ is the spectral effect of sound radiation at the lip opening and f is frequency in Hz. In the time domain, the effect of radiation at the lips is typically modelled as a first order differentiator, in which case Eq. (5.1) can be reduced to:

$$S(f) = G_{diff}(f)V(f) \tag{5.2}$$

where G_{diff} is the spectrum of the differentiated glottal flow. Thus, the glottal source derivative can be obtained by inverse filtering if the vocal tract transfer function is known.

However, $V(f)$ is not directly observable and as a result accurate estimation of the glottal source signal becomes an immensely difficult signal processing task (see e.g., Walker and Murphy, 2007; Alku, 2011). $V(f)$ is often treated as an all-pole model which can facilitate the use of Linear Predictive Coding (LPC) for estimating the parameters of the

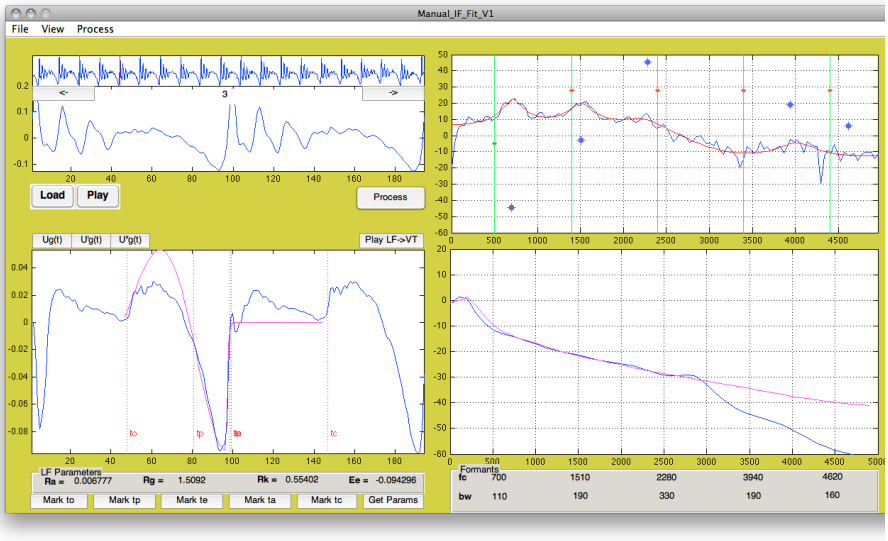


Figure 5.1: Screenshot of the in-house software for carrying out manually optimised inverse filtering and parameterisation.

vocal tract model. [1] Many automatic algorithms exist for vocal tract inverse filtering, including: closed-phase methods (Wong et al., 1979; Alku et al., 2009), iterative and adaptive Linear Predictive Coding (LPC) based methods (Alku, 1992), and methods which consider the mixed-phase properties of speech (Bozkurt et al., 2005; Drugman and Dutoit, 2009).

Due to the frequent problems of automatic algorithms, the research group has tended to rely on an inverse filtering method which derives initial estimates using an automatic closed-phase inverse filtering technique followed by an optimisation procedure involving manual fine-tuning (see the screenshot of the software in Figure 5.1). The user modifies the estimated formant frequencies and bandwidths and utilises both time and frequency domain displays to obtain maximum formant cancellation (see Chapter 2, Section 2.3.5 and Gobl and Ní Chasaide, 1999a).

The inverse filtering of the speech signal provides an estimate of the differentiated glottal source, which one can then characterise by fitting the Liljencrants and Fant (LF) source model (Fant et al., 1985a) to the individual glottal pulses, thus facilitating the parameterisation of important features in the source signal. Again, an automatic algorithm is first used to derive an initial model fit, which is followed by manual fine-tuning to get an improved fit. As with the inverse filtering, the user is visually guided to ensure optimisation in both the time and frequency domain. Furthermore, the user ensures that subsequent

[1] Note, however, that this is a simplification of the vocal tract system which does not consider the zeros which are often present in nasals

model pulses do not have unwarranted discontinuities. The manual fine-tuning involved in both the inverse filtering and the source parameterisation is extremely labour intensive. However, due to the limitations of current automatic algorithms, this approach has been deemed necessary if a precise description of the glottal source is required.

This chapter describes an automatic glottal source derivative parameterisation method which attempts to simulate some of the strategies used by the researcher when applying the visually guided optimisation of the model fit. To do this, an exhaustive search method is initially used which provides the N most suitable settings for the modelling, in terms of both time and frequency domain criteria. A dynamic programming algorithm is then used to select the best 'path' of parameter values by considering both the 'target cost' (i.e. the temporal and spectral match of the modelled pulses) and the 'transition cost' (i.e. the continuity in the parameter trajectories of the modelled glottal pulses). An optimisation method is then employed to refine the fit, again considering time and frequency domain criteria.

To evaluate the new method it's performance is compared to that of a standard automatic algorithm based on model fitting in the time domain as well as a parameterisation method based on phase minimisation. Reference values were obtained from manually optimised glottal source analysis. The effect of applying a widely used automatic inverse filtering algorithm was compared with a manual fine-tuning inverse filtering approach and demonstrate the effect on parameterisation. Furthermore, evaluation is carried out on a larger dataset where reference parameter values are obtained from simultaneous electroglottographic (EGG) signals.

5.2 State-of-the-art

Given an estimate of the glottal source derivative, a subsequent parameterisation stage is usually carried out in order to quantify the salient features of the signal. A recent study reviewed some of the more common methods typically used (Alku, 2011) and a review was also given in Chapter 2, Section 2.3.3.

This study uses the fitting method described in Strik et al. (1993) and Strik (1998) as a comparison as, like the proposed method, it operates on inverse filtered signals. Furthermore this approach is used in some current speech technology systems (see e.g., Cabral et al., 2011b). A further, more recently developed, comparison algorithm is also used (Degottex et al., 2011b) which estimates the shape of a glottal source model without the use of glottal inverse filtering (see Section 5.4.3).

5.3 Proposed method

A method is described here for estimating LF model parameter values (named **DyProg-LF**) by considering some of the information typically available to a researcher using a manual fine-tuning approach, i.e. time domain and frequency domain information, and overall parameter trajectory. It is assumed that the glottal source derivative has been derived beforehand using either an automatic method or using a manual fine-tuning method (as is used here).

5.3.1 LF model

The glottal source model used in the DyProg-LF method is the LF model. The model is described in Chapter 2, Section 2.3.2, with a more comprehensive description in Gobl (2003) and Fant et al. (1985a).

5.3.2 GCIs, f_0 and EE

In order to estimate f_0 and EE values from the glottal source derivative, the glottal closure instants (GCIs, which correspond to t_e in the LF model) are estimated using the SE-VQ algorithm described in Chapter 4. GCIs are shifted to the position of the maximum negative amplitude in the differentiated glottal source waveform, in the vicinity of the detected GCI. The maximum negative amplitude is used as the EE parameter value and f_0 is determined by the reciprocal of the duration between adjacent GCIs (in seconds).

5.3.3 Exhaustive search of Rd

Standard automatic time domain approaches to LF model fitting typically involve estimating initial parameter values by direct measurements of the differentiated glottal source pulse and then refining these estimates using an optimisation procedure. One common problem with this approach is that direct measurements can often yield poor initial parameter values. This is frequently due to inconsistency in marking the point of glottal opening, t_o (Alku et al., 2002). Subsequent use of an optimisation algorithm does little to rectify the problem. This is highlighted in Figure 5.2 where the final pulse from the time domain-based method (dot-dashed line) changes shape considerably from the previous pulses. In the second last pulse for the standard method the OQ value is around 0.5 and around 0.7 for the final pulse, whereas for the manual method it is approximately 0.5 for both pulses. An OQ of 0.5 is normally associated with a modal phonation type, with

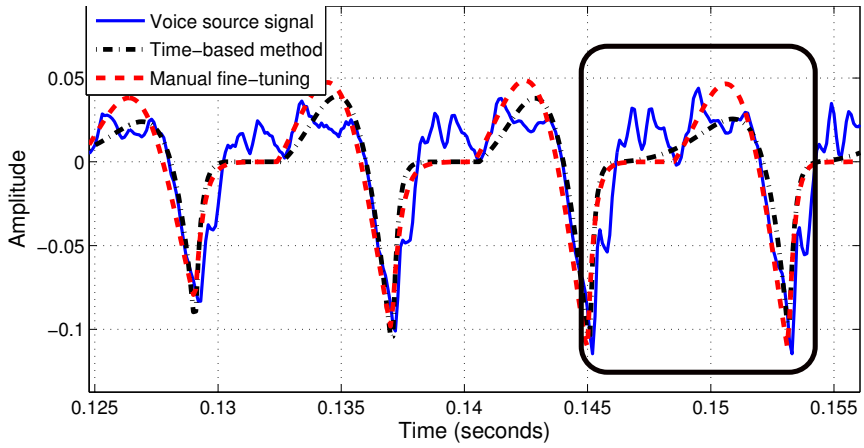


Figure 5.2: Estimated differentiated glottal source signal (solid blue line), a synthesised source signal using manually tuned parameters (red dashed line) and a source signal derived using parameters estimated using the standard automatic time based parameterisation method (black dot-dashed line). The change in the model setting in the final pulse highlights the potential for inconsistency.

0.7 indicating a laxer quality. As a result this type of sudden change could seriously affect the findings using this type of automatic parameterisation method.

To overcome this an exhaustive search method is proposed which involves the generation and analysis of a wide range of LF-model parameter configurations, and saving those configurations that minimise a specific error function.

However, to cover the full range of possible LF model configurations would have high computational load. Therefore, rather than searching all R-parameter combinations, the search is simplified by varying only Rd, and generating default Ra, Rk and Rg values from this (Fant et al., 1995, these equations are shown in Chapter 2, Section 2.3.2; Eqs. 2.17 - 2.19). Rd is changed in steps of 0.1 within the range [0.3, 3].

The search is done by first taking a GCI centred frame of the estimated glottal source derivative, $g'(n)$, and windowing it using a Hanning window. A frame length, L, of three local glottal periods is used to ensure clear harmonic peaks in the spectrum. The amplitude spectrum in dB is then calculated from the windowed glottal source derivative segment using the FFT. Harmonic amplitudes are measured in the spectrum, by searching the vicinity of integer multiples of the local f_0 up to a specific maximum frequency (H_{max}). Although methods exist for estimating the 'Maximum Voiced Frequency' (Stylianou, 2001; Zivanovic et al., 2007), these methods can at times lack robustness. As a result researchers often set a fixed H_{max} (Pantazis and Stylianou, 2008; Drugman et al., 2009c), and, hence, H_{max} is set to 3 kHz. For each step in the search an LF model pulse is generated using f_0

and EE (previously calculated) and using Ra, Rk and Rg as derived from the current Rd value. A synthetic signal is obtained by concatenating the LF pulses thereby producing a three-pulse segment, again centred on a GCI. The spectrum and harmonics are measured as above. For each Rd (and, hence, each generated three-pulse segment), an error value is measured between the two harmonic sets using:

$$\text{spec\_err} = \{1 - |cor\{h_U(m), h_{LF}(m)\}|\} \cdot w_s \quad 1 \leq m \leq N \quad \in [0, 1] \tag{5.3}$$

where h_U and h_{LF} are the harmonic amplitudes measured from the estimated glottal source derivative and the synthesised LF model signal respectively, N is the number of harmonics of frequencies below H_{max}, $cor\{.\}$ is the Pearson correlation between the harmonics h_U and h_{LF}, and w_s is a constant weight (see Section 5.4.2). A time domain error value is measured using:

$$\text{time\_err} = \{1 - |cor\{g_{LF}, g'(t)\}|\} \cdot w_t \quad \in [0, 1] \tag{5.4}$$

where g_{LF} is a synthesised LF model source signal of length L and set using the current Rd value, t is the sample range from the start and end point of the current frame, and w_t is a constant weight. Both time and frequency domain cost elements are designed to lie within the range [0, 1]. Another advantage of these correlation based error criteria is that, unlike more common error criteria (e.g., root-mean squared error), the value is independent of potential errors in EE.

The N_{cand} (empirically set to 5) Rd values are considered, that minimise the total error function:

$$\text{total\_cost} = \frac{\text{spec\_err} + \text{time\_err}}{2} \quad \in [0, 1] \tag{5.5}$$

5.3.4 Dynamic programming

A dynamic programming method is used to select the optimal path of Rd values through the input speech signal. The particular dynamic programming method used here is described in Ney (1983) and has been used in the popular *get_f0* pitch tracker (Talkin, 1995), as well as in the SE-VQ algorithm (Chapter 4).

The target cost, $d(i, j)$, is defined as the error value calculated in the exhaustive search (Eq. 5.5) for each Rd candidate in each analysis frame, where $1 \leq j \leq N_{cand}$, $1 \leq i \leq M$ and M is the number of GCIs (i.e. the number of analysis frames). The transition cost can be written as:

$$\delta_{i,j,k} = \{1 - cor\{seg_{i,j}, seg_{i-1,k}\}\} \cdot w_{tr} \cdot ss \quad \in [0,1] \tag{5.6}$$

where $seg_{i,j}$ refers to a single generated LF model pulse using the R-parameters predicted from the j-th Rd candidate at frame i and $seg_{i-1,k}$ refers to an LF pulse generated using k-th Rd candidate at the previous frame, $i-1$. This transition cost is based on the observation that, like the vocal tract, glottal source pulses should be reasonably slowly varying over a short timespan (e.g., 20 ms).

The transition cost is also dynamically modulated by the factor ss (the spectral stationarity measured used in Talkin, 1995):

$$ss = \frac{0.2}{\text{itakura}(f_i, f_{i-1}) - 0.8} \quad \in [0,1] \tag{5.7}$$

where itakura(.) is the Itakura distortion measure (Itakura, 1975)[2] of a glottal source derivative frame f_i (GCI centred, and with a duration three times the local glottal cycle) and the frame centred on the previous GCI, f_{i-1}. ss tends towards 1 when the spectral characteristics of adjacent frames are very similar and goes closer to 0 if the frames show a high degree of difference. This factor affects the transition cost, $\delta_{i,j,k}$, so that if there is an area of rapid change (e.g., at consonant-vowel transitions, at voice onset and offset, or for certain voice qualities involving irregular periodicity, such as creaky and harsh voice [3]) the transition cost has less of an effect. However, for relatively stable regions (e.g., the centre of vowels) the transition cost has a stronger effect in maintaining a smooth parameter contour. An example of the ss contour is shown in Figure 5.3 where ss declines at the voice offset around 1.02 seconds and remains low for the subsequent creaky voice region.

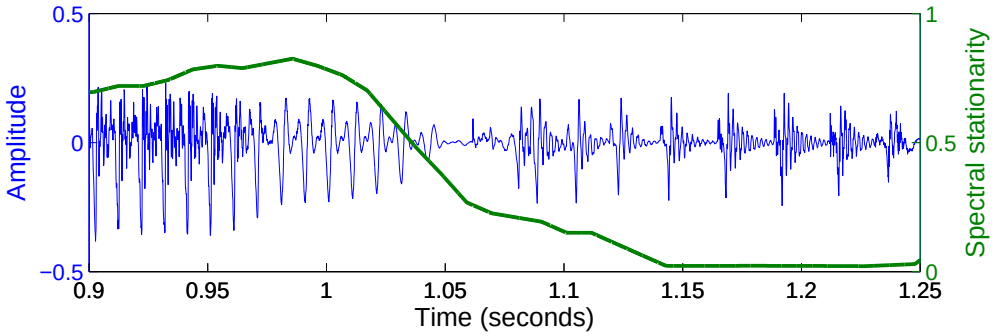


Figure 5.3: ss contour for sentence spoken by a male speaker. There is voicing offset from around 1.02 seconds and the region from around 1.08 to the end contains creaky voice.

[2]Note that the minimum Itakura distortion value is 1, and its upper bound value tends towards infinity.
[3]However, such voice qualities were not included in the data used in the current chapter

An objective function is, hence, defined for a given frame i incorporating the above target and transition costs (Eq. 5.5 and Eq. 5.6, respectively):

$$D_{i,j} = d_{i,j} + \min_{k \in N_{cand}} \{D_{i-1,k} + \delta_{i,j,k}\}, \quad 1 \leq j \leq N_{cand} \tag{5.8}$$

which is initialised with:

$$D_{o,j} = 0, \quad 1 \leq j \leq N_{cand} \tag{5.9}$$

The vector $q(i)$ is used to save the index of the optimal Rd (obtained by $\underset{j}{\text{argmin}}(D_{i,j})$ for $1 \leq i \leq M$).

An illustration of the optimal Rd path is shown in Figure 5.4 along with the N_{cand} Rd candidates at each frame (corresponding to each GCI). One can observe in the region from 0.9 to 1.0 seconds, just before the consonantal occlusion, that there is clearly rapid change in the signal and as a result Rd settings change quite considerably. This is due to the low spectral stationarity values in this region which severely lessens the effect of the transition cost. However, for the very stable regions (e.g., 0.7 to 0.85 seconds) the high spectral stationarity of successive frames ensures the transition cost has a strong effect and as a result very stable Rd values are observed.

5.3.5 Optimisation

Although the Rd parameter can be used to characterise many of the glottal pulse types arising in phonation types on a lax to tense continuum, it is likely that some glottal pulses will exist outside the constraints of Rd. Furthermore, it is not the intention to reduce the degrees of freedom of the model. To overcome this, parameter values are refined using an optimisation method. For each analysis frame Ra, Rk and Rg are derived from the Rd value, selected from the dynamic programming method. A simplex-based method (Nelder and Mead, 1965) is then used, which allows unbounded multi-variable optimisation. The three R-parameters are allowed to vary to minimise the same error function shown in Eq. (5.5).

5.4 Evaluation

In this section the procedure used to evaluate the DyProg-LF method is described. The experiments were designed to determine whether the method produces a better parameterisation of the glottal source than the comparison algorithms. Also investigated was the

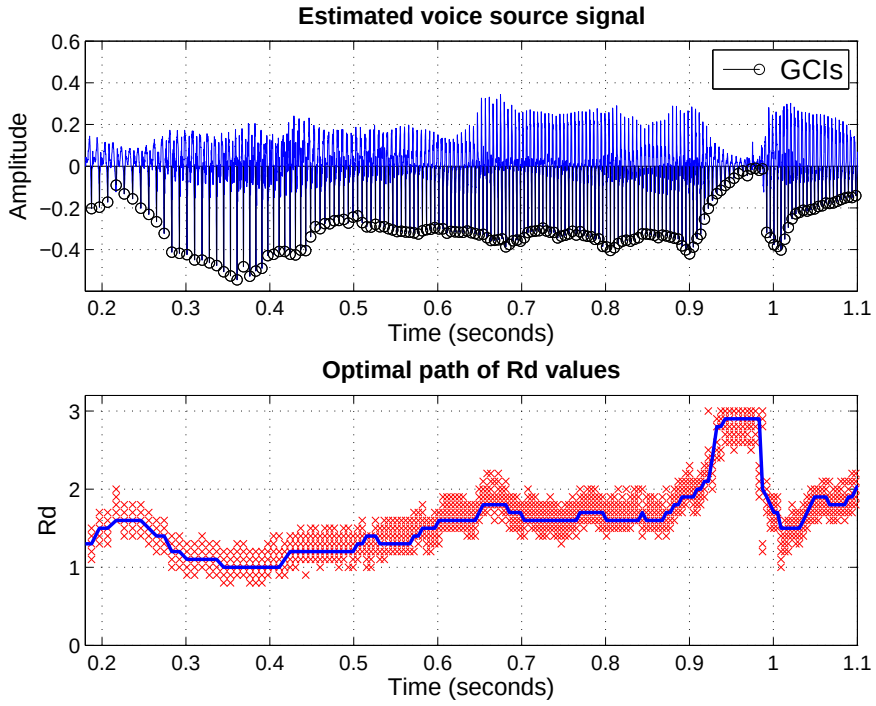


Figure 5.4: Voice source waveform estimated using the manual method and superimposed GCIs (top panel). Rd candidates (x's) and optimal Rd trajectory (line) plotted over time (bottom panel).

effect of manual versus automatic inverse filtering on the results of the parameterisation. Two types of evaluation were carried out; one objective and one qualitative.

5.4.1 Speech data

Speech data from six male speakers were used in the first part of the objective evaluation (see Section 5.4.5). Each speaker uttered the sentence *WE WERE aWAY a YEAR ago*, with narrow focus on each of the potentially accented syllables (*WE*, *WERE*, *-WAY* and *YEAR*) with both rising and falling pitch patterns. A broad focus and a deaccented rendition of the utterance were also recorded. Overall there were 10 utterances per speaker with the exception of one, from whom the 4 rising pitch utterances were not elicited. The speech samples for this speaker (6 utterances) were used for setting the weights (see Section 5.4.2) and were subsequently excluded from the testing, leaving 50 utterances for the first part of the objective evaluation.

For the qualitative evaluation (described in Section 5.4.7) a set of utterances produced by a male speaker was used (this is the same set of utterances as was analysed in Ní Chasaide et al., 2011). The sentence and the focus placements were the same as above,

but in this case only the narrow focus conditions was used.

Audio was captured in a semi-anechoic recording studio using high quality recording equipment (a B & K 4191 free-field microphone and a B & K 7749 pre-amplifier) and was digitised at 44.1 kHz (using a LYNX-two sound card), which was subsequently downsampled to 10 kHz. The DC-component and very low frequency components were removed using an 8th order high pass Butterworth filter with a cut-on frequency of 60 Hz. Filtering was carried out forwards and backwards to maintain the original phase spectrum of the signal.

For the second part of the objective evaluation (see Section 5.4.6) speech data from an American Female (SLT) and an American Male (BDL) speaker from the ARCTIC database (Kominek and Black, 2004) were used. The SLT set contained 1120 sentences and the BDL set contained 1130 sentences. These sentences involved a wide phonetic coverage and were recorded with simultaneous electroglottographic (EGG) signals. Both audio and EGG signals were originally sampled at 32 kHz, but were downsampled to 16 kHz for analysis.

5.4.2 Weight setting

In the evaluation, the DyProg-LF method is configured as described in Section 5.3. However, the weights w_s, w_t and w_{tr}, which are parameters of the dynamic programming method (Eq. 5.8), need to be set. The setting of these weights is crucial for modelling the relative importance of different types of information used in the manual fine-tuning approach and, hence, they need to be set carefully.

Using the speech data from one speaker (see Section 5.4.1) an exhaustive search was conducted to test all combinations of the three weights in the range [0, 1] with a step of 0.1 (1331 possible combinations). Note that all three cost elements were designed to lie within the range [0, 1]. For each combination, analysis was carried out on the 6 sentences and a synthetic glottal source signal was generated using the extracted parameter values. This was compared to a synthetic glottal source signal generated using the reference parameter values, by calculating the Pearson correlation coefficient (i.e. a measure of the similarity of the modelling by the automatic method compared to the modelling by the manual fine-tuning method was derived). The combination with the highest correlation score (averaged across the 6 sentences) was kept as the setting for the weights. The analysis resulted in the weights 0.6, 1 and 0.5 for w_s, w_t and w_{tr} respectively. This suggests that the manual fine-tuning user favours the time domain information for fitting the model, but frequency domain and continuity information also carry importance. These weights

appear to corroborate the author's subjective impression of the relative importance of these components.

5.4.3 Comparison algorithms

To evaluate the performance of the DyProg-LF method, two algorithms were selected from the literature to use as comparison. One algorithm is a traditional method for fitting LF model pulses to a given estimated glottal source derivative (Strik, 1998, see Chapter 2, Section 2.3.3 for the description). The second is a more recent development which estimates the *Rd* parameter of the LF model (Fant et al., 1995) by optimising a phase-based objective function (Degottex et al., 2011a, described in Chapter 2, Section 2.3.3).

5.4.4 Reference values

Objective evaluation of glottal source parameterisation is a difficult task. Some researchers tend to use synthetic stimuli to provide a quantitative evaluation with known reference parameter values (see e.g., Strik, 1998; Drugman et al., 2011). However these signals may lack the very details that cause trouble for glottal source parameterisation (e.g., the presence of aspiration noise). Others use EGG signals for obtaining reference values (Veeneman and BeMent, 1985; Henrich et al., 2001). It is not generally feasible, however, to obtain a full set of glottal source parameter values from the EGG signal. Other studies (e.g., Airas and Alku, 2007; Kane et al., 2010; Drugman et al., 2011) evaluate parameterisation on the basis of the ability of extracted parameters to differentiate voice quality. In the present chapter the DyProg-LF method is evaluated on two separate sets of speech data, using different methods for obtaining reference values in each.

In the first set, speech signals were inverse filtered by a member of the research group at the Phonetics and Speech Laboratory using the manual fine-tuning software (Gobl and Ní Chasaide, 1999a), where the user adjusts the formant frequencies and bandwidths for each analysis frame and uses time and frequency displays to achieve optimal formant cancellation.

With an estimated glottal source derivative, initial LF model settings were automatically derived for each glottal pulse. A member of the research group in the Phonetics and Speech laboratory, who is highly experienced with this type of analysis, then used the manual fine-tuning technique in order to ensure optimal fitting of the LF model to each glottal pulse of the glottal source derivative. These references values were obtained for a relatively small dataset, given the labour intensive nature of obtaining these values. Nonetheless, this approach provides a highly reliable means of evaluation and this set is

called 'carefully controlled data'. The reader can refer to Section 2.3.5 for a discussion of the criteria used in this process.

A second set was used to evaluate the performance of the DyProg-LF algorithm on larger volume of data containing wide phonetic coverage. As it was simply unfeasible to obtain manually derived reference values for this larger set of data, a single glottal source parameter, the open quotient (OQ), was instead derived from the simultaneous EGG signals available with this data. Glottal closure instants (GCIs) and glottal opening instants (GOIs) were derived from the EGG signal using the SIGMA method (Thomas and Naylor, 2009). This method involves applying a stationary wavelet transform and subsequent use of the group delay function in order to derive candidate values. Candidate selection is carried out using a K-means algorithm, which is used to separate the candidates into two clusters based on a three dimensional feature vector.. The glottal open duration was calculated from GCIs and GOIs and normalised to the local glottal period to derive reference OQ values.

5.4.5 Objective evaluation - Part 1: carefully controlled data

Using the first set of reference values, the performance of the DyProg-LF method was evaluated by comparing it to the performance of the comparison methods (described in Section 5.4.3). The four R-parameters, Rg, Rk, Ra and Rd were considered and relative error was used as an evaluation metric:

$$\text{Relative error} = \frac{|param_{ref} - param_{est}|}{param_{ref}} \tag{5.10}$$

where $param_{ref}$ are the reference parameter values and $param_{est}$ are the parameter values as estimated by the automatic algorithms. Note that as the Degott-LF method only estimates the Rd parameter, comparison between all three parameterisation algorithms was only carried out for this parameter. For the other three parameters only the Strik-LF method was compared to the proposed algorithm.

As another goal of the evaluation is to assess the effect of automatic versus manual inverse filtering, the above experiments were carried out both on glottal source signals obtained using the manual fine-tuning approach and also on glottal source signals automatically derived using the Iterative and Adaptive Inverse Filtering (IAIF) method (Alku, 1992). A previous study found complex cepstrum-based decomposition (Drugman et al., 2009a) and closed-phase inverse filtering (Yegnanarayana and Veldhuis, 1998) to outperform IAIF in certain experiments (Drugman et al., 2011). However, the complex-cepstrum

method does not separate the glottal return phase from the vocal tract component, an important aspect of the glottal pulse which the author wishes to consider in this study. This is because both the vocal tract and glottal return phase are contained with the minimum-phase component of the speech frame. Closed-phase inverse filtering is also very sensitive to slight errors in the positioning of the glottal closure and glottal opening instants. In fact, experimentation with IAIF by the author (see Chapter 6) has suggested much more comparable performance to these methods than has previously been reported. Furthermore, recent speech synthesis methods utilising glottal inverse filtering have opted to use IAIF over other methods (Cabral et al., 2011b; Raitio et al., 2011).

Statistical analysis

The aim of the statistical analysis was to test the effect of parameterisation algorithm type and inverse filtering method on the evaluation metrics and to generalise over both speakers and utterances. Linear mixed-effects modelling (Baayen, 2008) facilitates doing this in one single analysis rather than conducting two separate ANOVAs (in order to generalise over both speakers and utterances), and, hence, is used in the current study. This type of statistical analysis has become popular in speech analysis papers in recent years as an elegant method for modelling fixed and random effects in combination (used for example in a study by Vainio et al., 2010 on a similar topic).

Both parameterisation algorithm type and inverse filtering method are treated as fixed effects, with speaker and utterance treated as random effects and with each evaluation metric considered as the dependent variable. p-values were obtained using Monte Carlo Markov Chain (MCMC) sampling (using 10,000 samples).

The statistical analysis was carried out using the R statistical software platform using the `lme4` package developed by Baayen (2008). The linear mixed effects modelling was carried out using the following command (here exemplified):

```
Rg_err.lme <- lmer(Rg_err~param * IF+(1|sentence)+(1|speaker),data=timeFreq_data)
```

where `param` is the variable name for parameterisation algorithm type factor and `IF` is the variable name for the inverse filtering type factor.

Estimation of p-values using MCMC sampling is done using:

```
Rg_err.p <- pvals.fnc(Rg_err.lme)
```

Note that this procedure was carried out only for *Ra*, *Rk* and *Rg*. As the Degott-LF method does not involve glottal inverse filtering, the parameterisation and inverse filtering groups were merged for analysis of the *Rd* parameter. This gave 5 methods: Degott-LF,

proposed (DyProg-LF) with manual inverse filtering, proposed with IAIF, Strik-LF with manual inverse filtering and Strik-LF with IAIF. So, for the *Rd* parameter there was only a single fixed effect.

5.4.6 Objective evaluation - Part 2: large database analysis

For the second part of the objective evaluation using the larger dataset absolute error was used as the evaluation metric, as the *OQ* parameter is bound in the range [0, 1]. For each sentence in the SLT and BDL datasets absolute error values were calculated for the three algorithms. The median of these error values, for each method, was calculated for every sentence and these were used as datapoints in the subsequent statistical analysis.

Note that for the DyProg-LF method and for Strik-LF, *OQ* can easily be derived from a given LF model fit using Eq. (2.20). For Degott-LF, however, as it only estimates the *Rd* parameter, the conversion to *OQ* is less straightforward. In this study *OQ* was derived from *Rd* for Degott-LF by exploiting previous regression analysis reported in (Fant et al., 1995), using the equations in Chapter 2, Section 2.3.2. Rg_p and Rk_p are derived from Eqs. 2.19 and 2.18, respectively. *OQ* is derived using Rg_p and Rk_p as inputs to Eq. (2.20). For this process the *EE* value is measured as the strongest negative peak in the estimated glottal source derivative, in the vicinity of the estimated *Rd* position. Although this is less straightforward for comparing performance with Degott-LF, the fact that *OQ* is the only LF model-based parameter which can be estimated from the EGG signal necessitates this procedure.

Statistical analysis

To investigate whether there were significant differences between the *OQ* values derived using the DyProg-LF method and the comparison methods a one-way ANOVA was applied with *OQ* absolute error as the dependent variable and algorithm type as the independent variable, for the SLT and BDL databases separately. Pairwise comparisons were carried out using Tukey's Honestly Significant Difference (HSD) test.

5.4.7 Qualitative evaluation

Although the objective evaluation described above provides a strong test of the performance of the method, a qualitative examination of the extracted parameter contours will provide complementary evidence of its performance.

A recent study by Ní Chasaide et al. (2011) presented an analysis of glottal source dynamics in utterances with varying focal placement (see Section 5.4.1). The glottal source data provided evidence for a tenser mode of phonation in focused syllables. One of the parameters analysed, the open quotient (OQ, see Eq. 2.20), has been shown to be effective in differentiating breathy and tense voice (Hanson, 1997; Airas and Alku, 2007) and for the qualitative evaluation this parameter will be focused on.

The evaluation involves examining the contours for OQ, as extracted by the manual method, the DyProg-LF method and the standard time-domain method (Strik-LF), and analysing to what extent the results presented in Ní Chasaide et al. (2011) are replicated when using automatic algorithms. Note that the Degott-LF method was excluded from this part of the evaluation.

5.5 Results

5.5.1 Objective evaluation

The distributions of relative error scores, between reference and estimated parameter values, are presented in Figures 5.5 - 5.8, with a summary of the statistical analysis for the three parameters, Ra, Rk and Rg presented in Table 5.2.

For Rg the boxplots in Figure 5.5 demonstrate a clearly lower relative error for the proposed parameterisation method (DyProg-LF) for both automatically and manually inverse filtered signals. This observation is supported by the statistical analysis and shown to be significant [t = 7.521, $p_{MCMC} = 0.0001$]. Another important observation is that the variance in relative error values is considerably larger for the standard parameterisation method compared to DyProg-LF. Interestingly a comparison of the distributions of relative error values for automatic and manual inverse filtering methods shows no significant differences [t = 0.173, $p_{MCMC} = 0.8734$].

A similar trend can be observed for the Rk parameter (see Figure 5.6) with the distribution of relative error values for the DyProg-LF method being significantly lower [t = 5.696, $p_{MCMC} = 0.0001$] than those for the time domain method. Again the inverse filtering type is found not to have a significant effect [t = -1.016, $p_{MCMC} = 0.3216$]. The size of the variance in the manual inverse filtering condition is found to be similar across the two parameterisation algorithms; however when automatic inverse filtering is used the standard parameterisation method produces a considerably larger variance in relative error scores.

Relative error scores are generally higher for Ra compared to the other two parameters

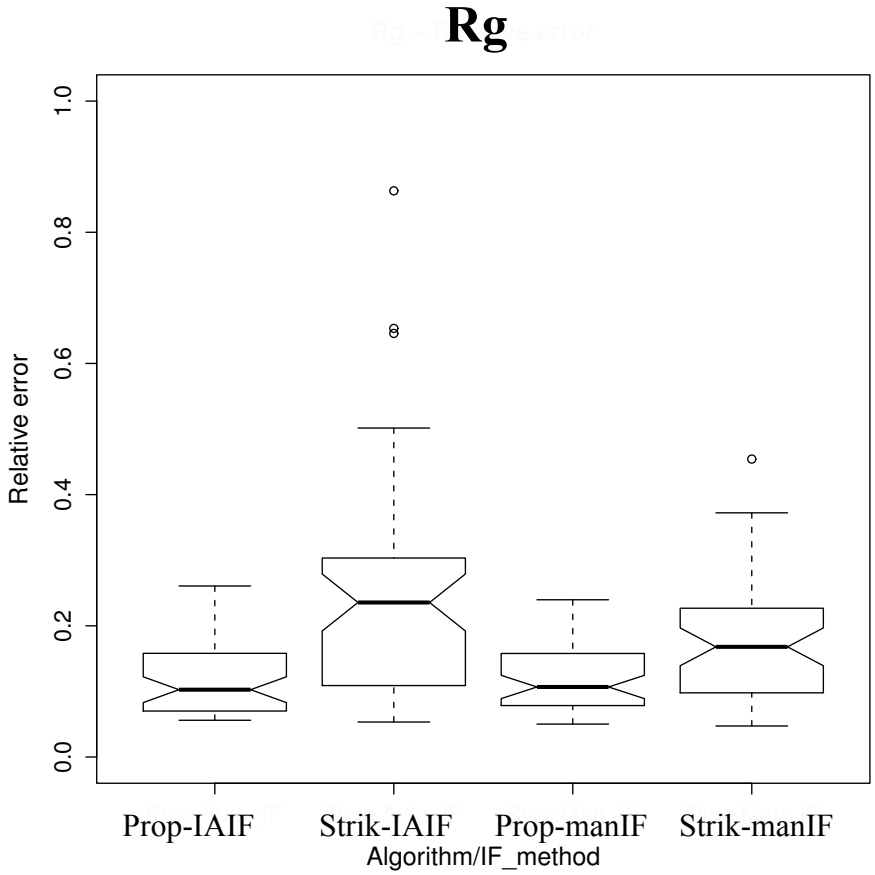


Figure 5.5: Distributions of relative error scores for **Rg** as a function of the four possible combinations of parameterisation algorithm type and inverse filtering type. **Prop** indicates proposed parameterisation method (DyProg-LF) with **Strik** indicating the standard time domain method. **IAIF** stands for automatic iterative adaptive inverse filtering with **manIF** being the manual inverse filtering method.

(see Figure 5.7). Here automatic inverse filtering has a very clear negative impact on relative error scores compared to manual inverse filtering [t = -5.820, $p_{MCMC} = 0.0001$]. Despite the strong effect of inverse filtering method, the improvement in relative error scores for the DyProg-LF method still achieves significance levels [t = 2.579, $p_{MCMC} = 0.0094$]. It is clear from the boxplot in Figure 5.7 that automatic inverse filtering severely affects the relative error score for the DyProg-LF method. However, the mean error is still lower than the mean error for the standard method in both inverse filtering conditions.

Relative errors are shown for *Rd* in Figure 5.8. Statistical analysis revealed no significant effect of analysis method on the relative error score [t = -1.839, $p_{MCMC} = 0.072$]. Despite this DyProg-LF combined with manual inverse filtering provided the lowest mean error scores. However, the Strik-LF method provided much closer relative error scores

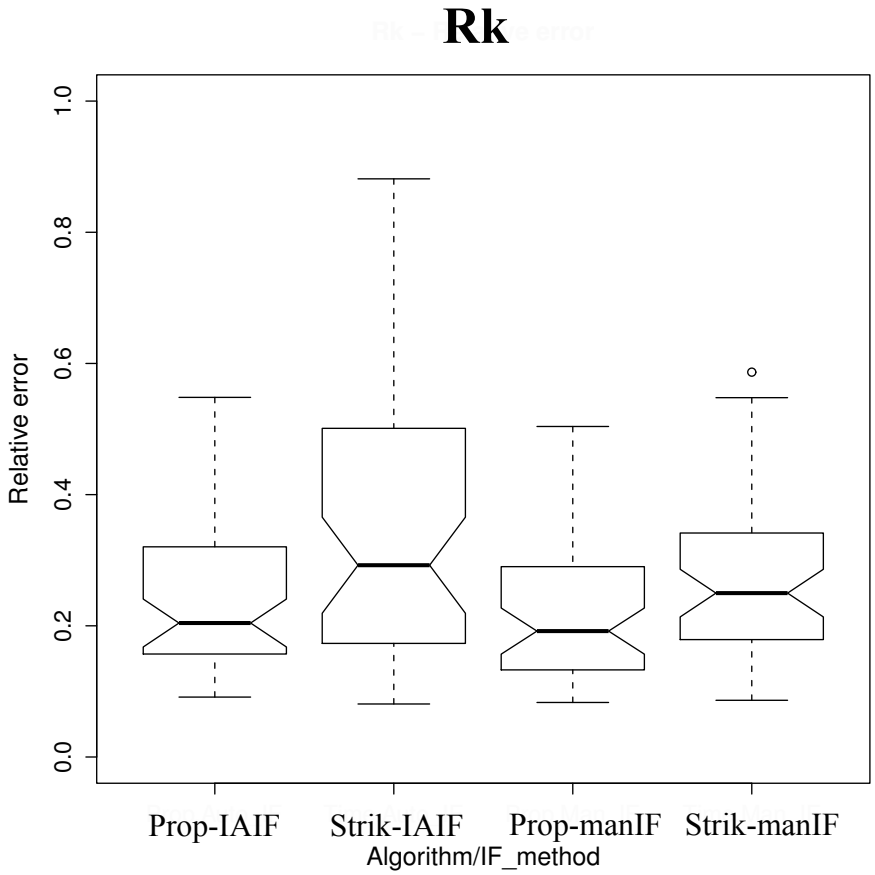


Figure 5.6: Distributions of relative error scores for **Rk** as a function of the four possible combinations of parameterisation algorithm type and inverse filtering type. **Prop** indicates proposed parameterisation method (DyProg-LF) with **Strik** indicating the standard time domain method. **IAIF** stands for automatic iterative adaptive inverse filtering with **manIF** being the manual inverse filtering method.

to the DyProg-LF method than for the other three *Rd* parameters. Degott-LF also gave similar mean relative error scores, but with higher levels of variance than for the other methods.

Results from the second part of the objective evaluation are illustrated in Figure 5.9, with the distributions of *OQ* errors plotted as a function of algorithm type for the SLT (left panel) and BDL (right panel) databases. ANOVAs revealed significant differences between the algorithms for both SLT [$F_{(2,3357)} = 7403.1$, $p < 0.001$] and BDL [$F_{(2,3387)} = 1625.8$, $p < 0.001$]. Posthoc testing using Tukey's HSD demonstrated highly significant differences ($p < 0.001$) for all pairwise comparisons for both SLT and BDL datasets. Considering Figure 5.9 it is clear that the DyProg-LF method produces the lowest *OQ* error compared to the two other algorithms. Strik-LF produced relatively high error scores with a very large

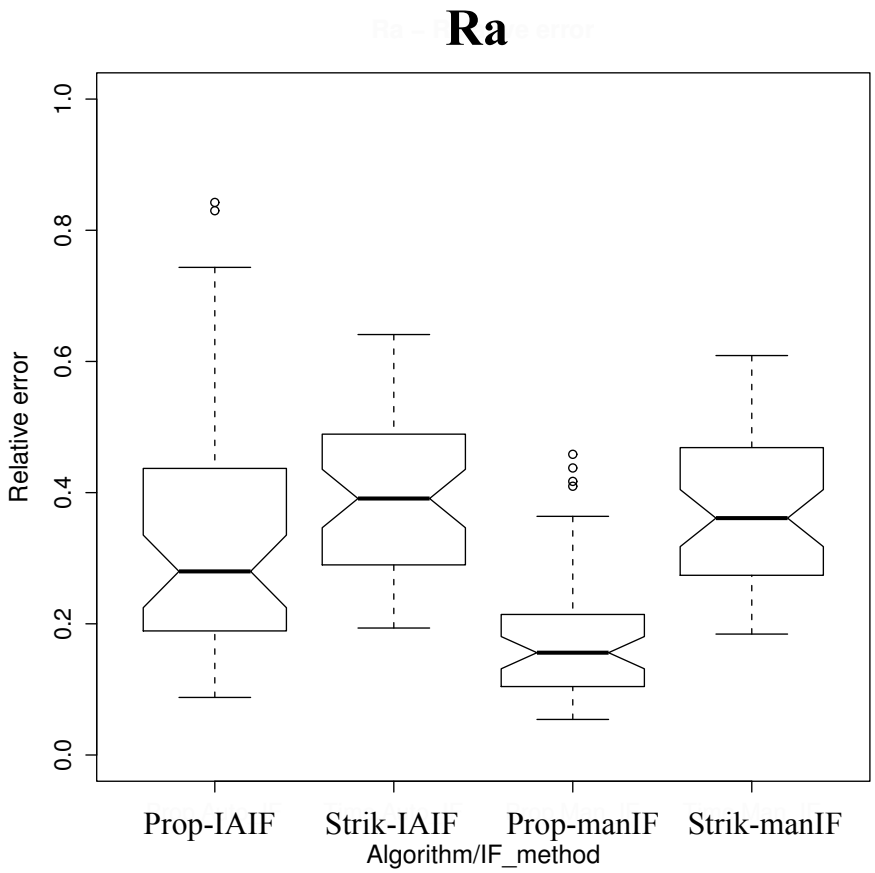


Figure 5.7: Distributions of relative error scores for **Ra** as a function of the four possible combinations of parameterisation algorithm type and inverse filtering type. **Prop** indicates proposed parameterisation method (DyProg-LF) with **Strik** indicating the standard time domain method. **IAIF** stands for automatic iterative adaptive inverse filtering with **manIF** being the manual inverse filtering method.

variance for BDL. Overall error scores were lower for the male speaker (BDL) than the female speaker (SLT). The higher f_0 in the female voice may have reduced the effectiveness of the glottal inverse filtering, resulting in the estimated glottal pulses not displaying a clear closed phase. This would have negatively affected the performance of both the DyProg-LF and Strik-LF algorithms. The method Degott-LF produces significantly lower ($p < 0.001$) OQ errors than the Strik-LF method for both BDL and SLT databases.

5.5.2 Qualitative evaluation

The *OQ* contours extracted using the manual method and using the DyProg-LF and the Strik-LF automatic algorithms are shown, along with the corresponding f_0 contours, for

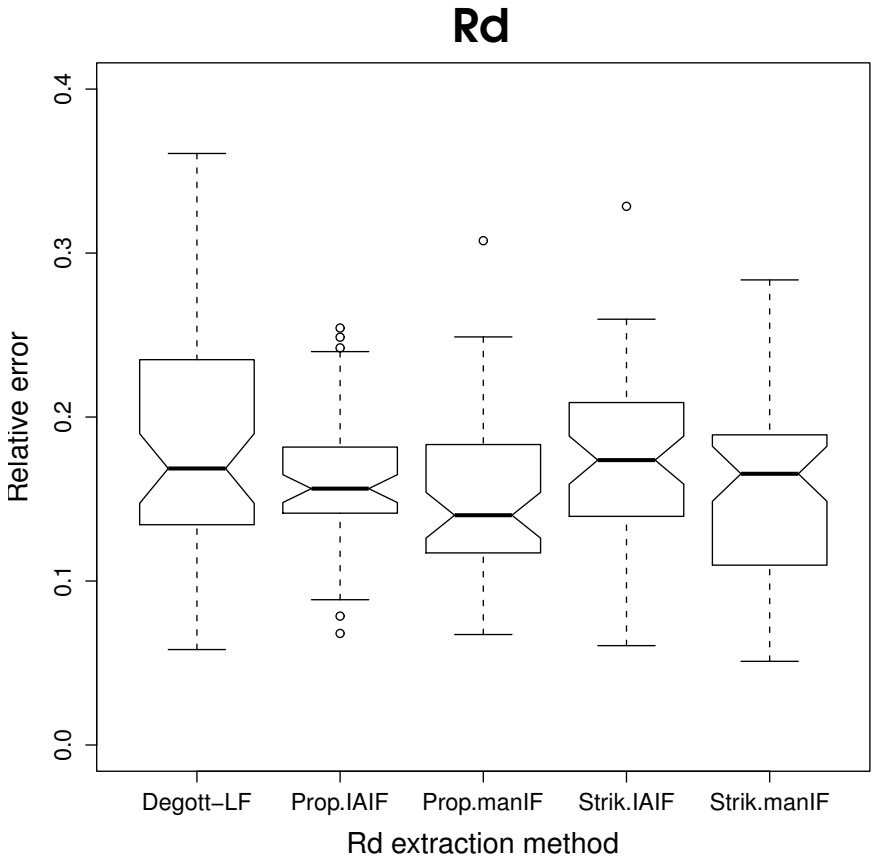


Figure 5.8: Distributions of relative error scores for **Rd** as a function of the five different means of deriving Rd values. **Prop** indicates proposed parameterisation method (DyProg-LF) with **Strik** indicating the standard time domain method. **IAIF** stands for automatic iterative adaptive inverse filtering with **Man_IF** being the manual inverse filtering method. **Degott-LF** indicates parameter estimation using the phase minimisation method.

falling intonation in Figures 5.10 - 5.11 and for rising intonation in Figures 5.12 - 5.13. Both figures have four separate panels for each of the four focused syllables and each panel is segmented with dashed lines and phonetic transcriptions are shown at the bottom of each panel. In Ní Chasaide et al. (2011) the glottal source data indicated a tenser phonation in focused syllables, and as low OQ values also point to this, the minimum value in each contour is highlighted with a star. Also, a summary of the size of the error values on OQ for the two automatic methods compared to references values is given in Table 5.3.

It can be seen in Figure 5.10, panels (a) and (b), that the minimum OQ value for the time-domain based method does not fall on the focused syllable, whereas minimum OQ values from the manual and proposed methods both fall on the focused syllable. When

Table 5.1: Summary of the results from the statistical analysis involving linear mixed effects modelling and subsequent p-value estimation using Monte Carlo Markov Chain sampling. Results shown are t and p_{MCMC} values for Rd.

Parameter	**Effects**	t**-value**	p_{mcmc} **value**
Rd	Method	-1.839	0.071

Table 5.2: Summary of the results from the statistical analysis involving linear mixed effects modelling and subsequent p-value estimation using Monte Carlo Markov Chain sampling. Results shown are t and p_{MCMC} values for the three **R-parameters** considering each of the fixed effects individually as well as the interaction of the two.

Parameter	**Effects**	t**-value**	p_{mcmc} **value**
Rg	Parameterisation	7.521	0.0001
	Inverse filtering	0.173	0.8734
	Interaction	-2.783	0.0066
Rk	Parameterisation	5.696	0.0001
	Inverse filtering	-1.016	0.3216
	Interaction	-2.337	0.0204
Ra	Parameterisation	2.579	0.0094
	Inverse filtering	-5.820	0.0001
	Interaction	3.470	0.0002

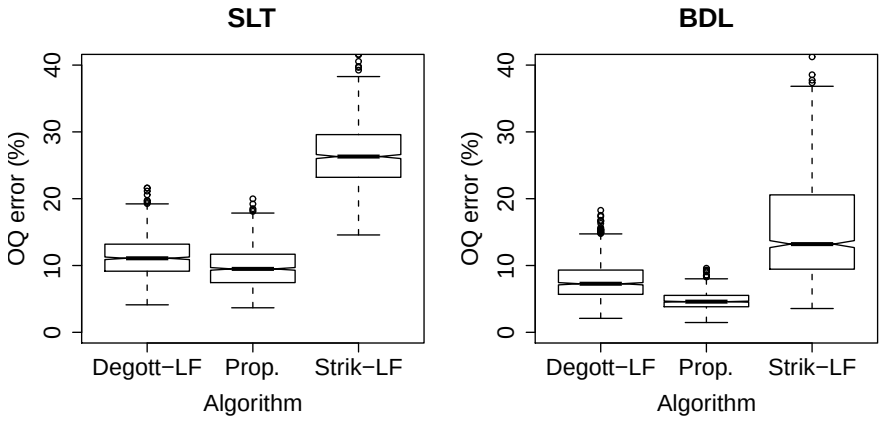


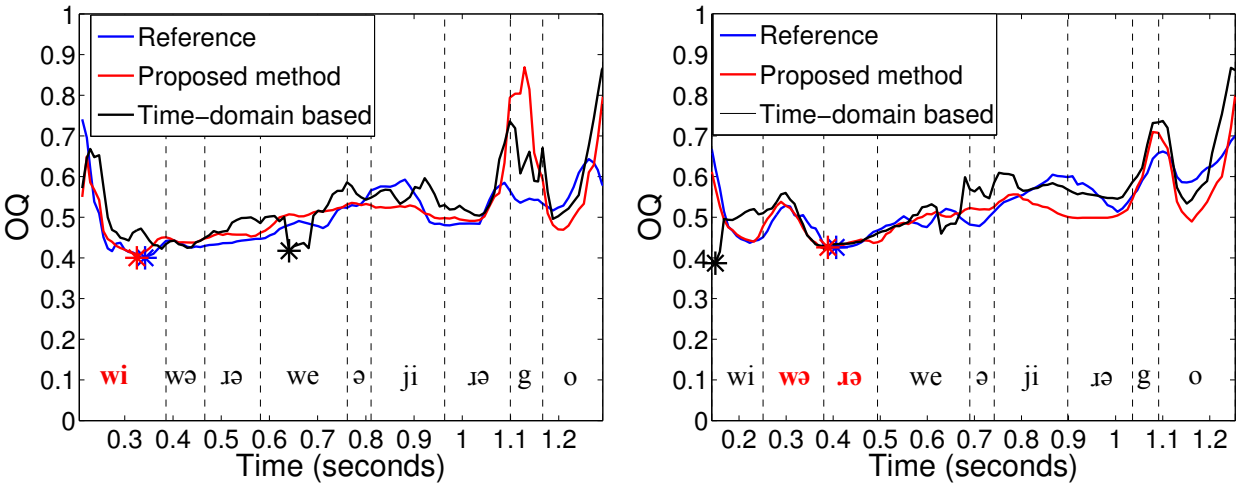
Figure 5.9: Distributions of absolute error (%) for OQ values derived using the three parameterisation methods compared to reference OQ values measured from simultaneous EGG signals, for the SLT (left panel) and BDL (right panel) ARCTIC databases.

Table 5.3: Mean ($\bar{x}$) and standard deviation (σ) of the absolute error score on OQ compared with the references values for the proposed method (DyProg-LF) and the method by Strik. Values are given for all four focused syllable conditions in both falling and rising intonation patterns.

Intonation patterns	**Focused syllable**	PROPOSED		STRIK	
		$\bar{x}$ (%)	σ (%)	$\bar{x}$ (%)	σ (%)
Falling	WE	2.5	2.3	3.6	2.3
	WERE	3.1	2.9	3.5	2.8
	-WAY	1.7	1.2	4.0	8.8
	YEAR	2.3	1.4	2.6	2.0
Rising	WE	4.4	4.2	9.2	10.0
	WERE	2.9	2.9	4.4	7.1
	-WAY	2.6	2.1	6.5	5.7
	YEAR	3.6	2.9	12.6	13.2

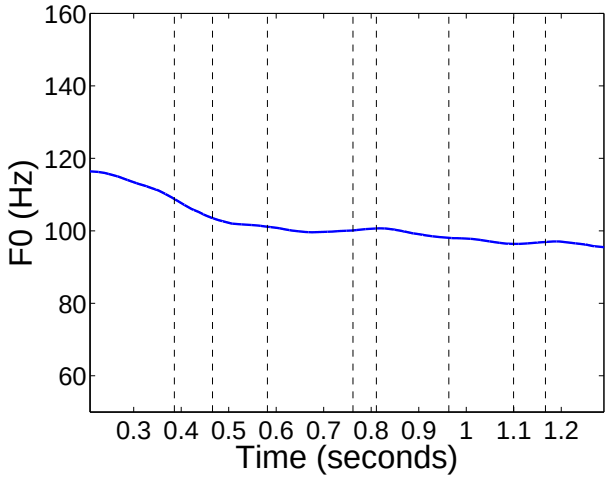
-WAY is the focused syllable (panel c) all three methods display minimum *OQ* values on the focused syllable. When -YEAR is focused, however, the two automatic algorithms both produce minimum *OQ* values after the minimum value from the manual method. Considering the error scores for the two methods in Table 5.3 one can observe a slight improvement for the DyProg-LF method.

For rising intonation patterns (shown in Figure 5.12), however, *OQ* contours for the standard time-domain based method (Strik-LF) are considerably more erratic, particularly for the focused syllables WE (panel a) and YEAR (panel d). The *OQ* contour from the proposed method (DyProg-LF) generally follows the reference *OQ* values more closely. When WERE (panel b) and -WAY (panel c) are in focus all three methods produce a minimum *OQ* value on the focused syllable. However, for the focused WE (panel a) only the reference values and the DyProg-LF method follow this trend. For the utterance where YEAR is the focused syllable (panel d) both the reference method and the time domain method both have minimum *OQ* values that fall on /we/. The minimum *OQ* value for the DyProg-LF method actually falls on the focused syllable. It can be observed that the reference *OQ* values also dip to relatively low values here, whereas the time domain *OQ* values rise around this time location.

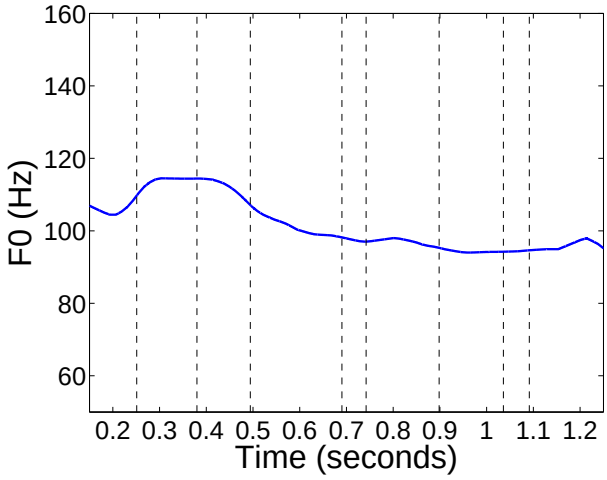


(a) WE (OQ)

(b) WERE (OQ)

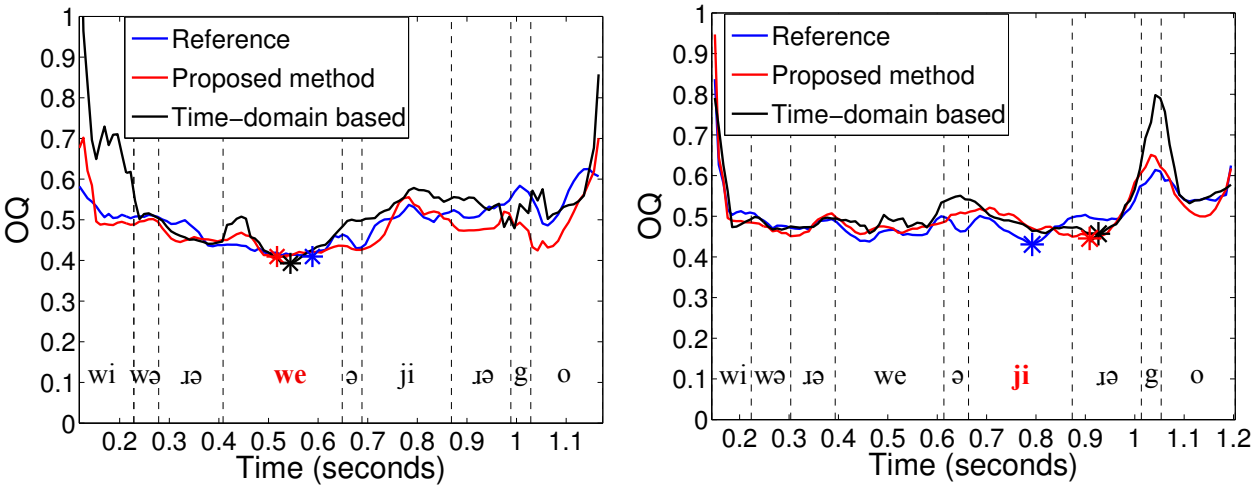


(c) WE (f_0)



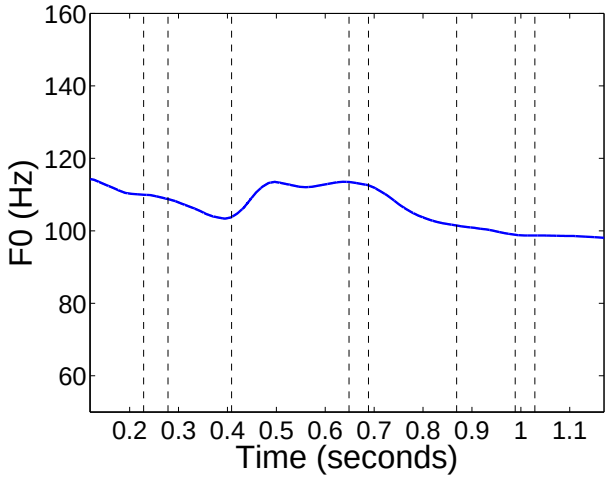
(d) WERE (f_0)

Figure 5.10: OQ contours for the reference manual method (blue), the proposed method (red) and the time domain based method (black), on utterances with **falling** intonation and narrow focus on the syllable we (a) and were (b), with corresponding f_0 contours, panels c and d, respectively. The whole sentence was *We were away a year ago* read by a male speaker. Parameter contours are smoothed with a 5-point moving average filter.

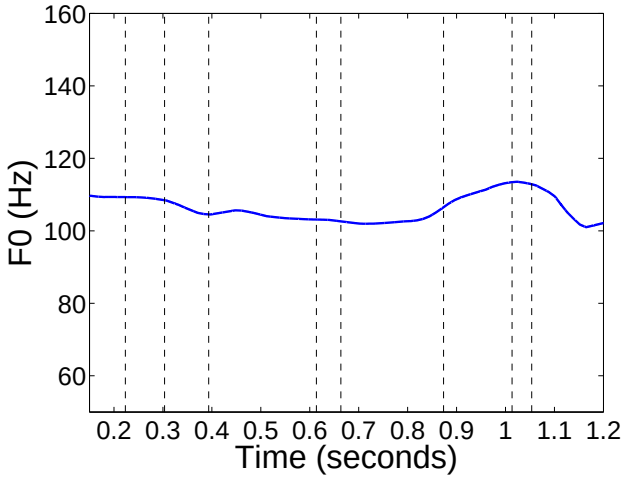


(a) -WAY (OQ)

(b) YEAR (OQ)

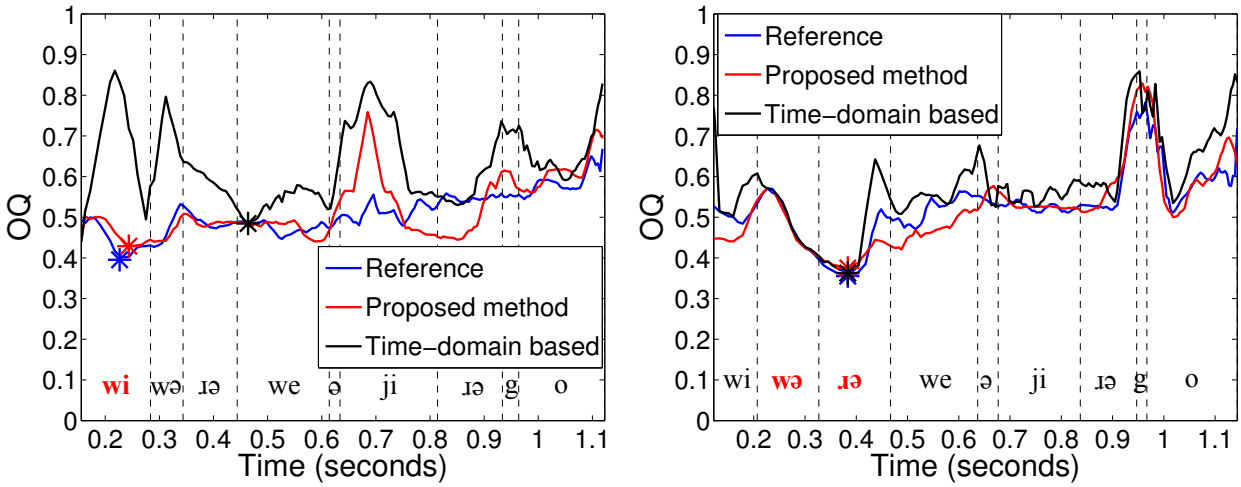


(c) -WAY (f_0)



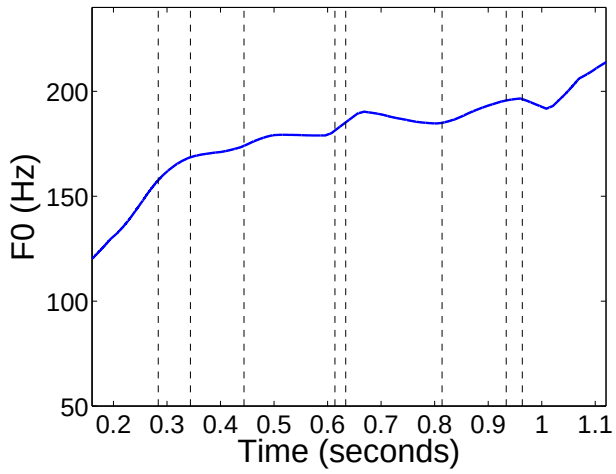
(d) YEAR (f_0)

Figure 5.11: OQ contours for the reference manual method (blue), the proposed method (red) and the time domain based method (black), on utterances with **falling** intonation and narrow focus on the syllable way (a) and year (b), with corresponding f_0 contours, panels c and d, respectively. The whole sentence was *We were away a year ago* read by a male speaker. Parameter contours are smoothed with a 5-point moving average filter.

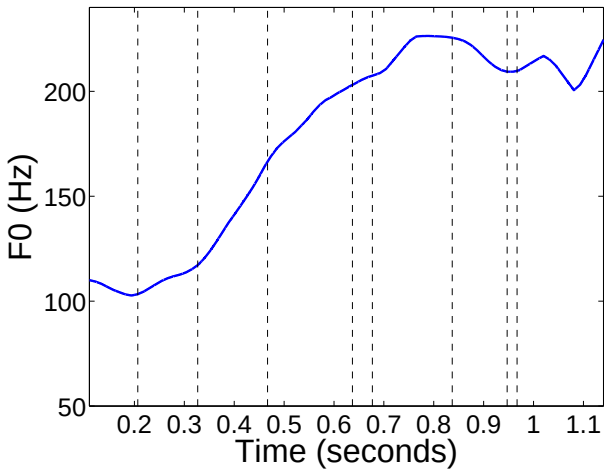


(a) WE (OQ)

(b) WERE (OQ)

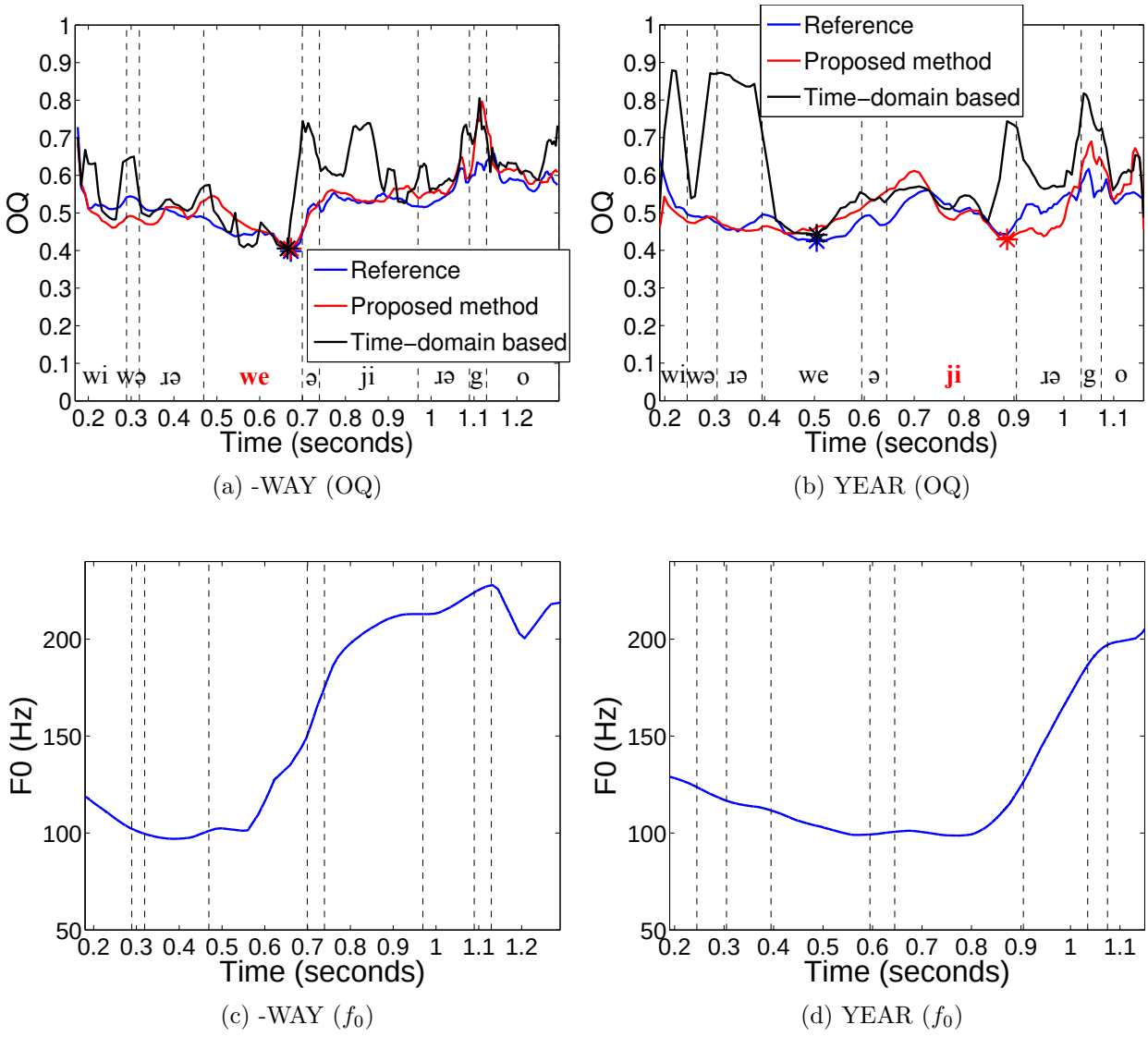


(c) WE (f_0)



(d) WERE (F_0)

Figure 5.12: OQ contours for the reference manual method (blue), the proposed method (red) and the time domain based method (black), on utterances with **rising** intonation and narrow focus on the syllable we (a) and were (b), with corresponding f_0 contours, panels c and d, respectively. The whole sentence was *We were away a year ago* read by a male speaker. Parameter contours are smoothed with a 5-point moving average filter.



(a) -WAY (OQ)

(b) YEAR (OQ)

(c) -WAY (f_0)

(d) YEAR (f_0)

Figure 5.13: OQ contours for the reference manual method (blue), the proposed method (red) and the time domain based method (black), on utterances with **rising** intonation and narrow focus on the syllable we (a), were (b), -way (c) and year (d). The whole sentence was *We were away a year ago* read by a male speaker and phonetic transcription is shown at the bottom of each panel. Parameter contours are smoothed with a 5-point moving average filter.

5.6 Discussion

The objective evaluation in this chapter revealed parameter values for the proposed method, called DyProg-LF, to be clearly closer to the references values compared to those for the standard time-domain based method (Strik-LF, Strik, 1998) for the parameters Ra, Rk and Rg. This is likely due to inconsistencies in determining the point of glottal opening, t_o, using the standard method (Alku et al., 2002). A combination of the exhaustive search and the dynamic programming components appear to be more suited to handling this problem. Previous application of this dynamic programming algorithm in f_0 and formant tracking (Talkin, 1995), as well as for GCI detection (Chapter 4), has already shown this approach to be useful for obtaining smooth parameter contours.

For Rd, however, the difference between the algorithms is less clear. This is likely due to the Strik-LF algorithm fitting to the amplitude of glottal source pulses consistently even if there is inconsistency in the marking of t_o. It is, hence, highlighted here that focusing on Rd alone would not provide this distinction.

Analysis of the larger corpora served to demonstrate the usefulness of the proposed method for carrying out large-scale glottal source modelling. The OQ parameter here demonstrates the inconsistency of Strik-LF for this purpose. The more recent, phase minimisation approach (Degottex et al., 2011b) is clearly more robust for this type of analysis. Although the proposed method produced lower errors in OQ estimation, this finding must be treated as tentative, given that the phase minimisation approach does not directly output OQ values and a prediction of OQ values from previous regression analysis (Fant et al., 1995) needed to be applied to allow comparison.

A further interesting finding from the quantitative evaluation was that for the Rg and Rk, parameters extracted using the DyProg-LF method automatic inverse filtering gave very similar error scores compared to manual inverse filtering. This, however, was not the case for the standard parameterisation method which produced higher errors when automatic inverse filtering was used. For the Ra parameter, inverse filtering type clearly had a big impact on the error scores for the proposed method. This is not surprising as one would expect the higher end of the spectrum to be relatively more sensitive to different inverse filter settings. This finding suggests that a combination of the IAIF inverse filtering (Alku, 1992) with the DyProg-LF method may provide a fully automatic way for extracting reliable open phase glottal source characteristics for male speakers.

The qualitative evaluation demonstrated that for the one male speaker dataset that the DyProg-LF method produces OQ contours that are very similar to those obtained using the manual fine-tuning approach. The findings of glottal source parameters pointing to a

tenser voice quality in focused syllables in a previous study (Ní Chasaide et al., 2011) would also have been arrived at if the DyProg-LF method had been used for parameterisation. The standard parameterisation method, however, did not corroborate these results as closely.

Although the qualitative evaluation only looked at one speaker, the results in combination with those from the objective evaluation support the notion that the proposed method could be used to help alleviate the workload of the manual fine-tuning approach, yet still maintain the precision. Much of the research attention in terms of the development of automatic glottal source analysis tools has focused on the inverse filtering component (e.g., Alku, 1992; Alku et al., 2009). However, it is quite clear from the present study that the parameterisation method used is also crucial in determining characteristics of the glottal source.

In terms of the proposed algorithm itself it was found that the combination of the exhaustive search and dynamic programming components provided a strong estimation of the open-phase parameters (i.e. Rg and Rk). As a result the subsequent optimisation component typically involved only a slight modification to these parameters. However, for Ra the optimisation component often involved a considerable change to the Ra setting, predicted from the Rd value (using the results of the regression analysis in Fant et al., 1995). As the prediction of the open-phase parameters was more robust, this suggests that refinements may be required to the prediction method for Ra. Furthermore, as the regression analysis carried out in Fant et al. (1995) was on a limited number of speakers, it may be that certain speakers deviate to a greater or lesser extent from this prediction model, particularly with respect to Ra. Furthermore, the second figure in Fant et al. (1994) shows some instances which deviate quite substantially from the regression-based prediction used. A future study, involving a range of speakers and with substantial variation in voice quality would be very beneficial for examining and improving the robustness of regression prediction using Rd described in Fant et al. (1995).

Another point on Ra is that it can be rather difficult to get an optimal fit to the return phase by only considering time domain information. Variation in Ra can bring rather small changes to the return phase in the time domain, but can have a rather profound effect on the spectral tilt in the frequency domain. The inclusion of frequency domain measurements in the error criterion used in DyProg-LF may help provide simultaneously better modelling of both the return phase and the higher frequencies in the glottal source derivative spectrum. At the same time, however, resolving open phase glottal source parameters solely from frequency domain measurements can be problematic (Henrich et al., 2001; Doval et al., 2006; Vincent, 2007; Degottex et al., 2011b). In particular, the com-

bined effect of Rg and Rk (and similarly OQ and the asymmetry parameter, α_m) have a complex effect on the lower end of glottal source spectrum, making it extremely difficult to derive the two separately from spectral measurements. Both these observations may have contributed to the weighting of the frequency domain error criterion (i.e. 0.6) being lower than the time domain error weighting (i.e. 1). Although the weighting is lower, it is still found to have an important contribution.

Findings from the first part of the objective evaluation must still be considered tentative considering that there was only a limited number of speakers analysed in and that the utterances produced were restricted to a particular set. Nevertheless, the present author believes that this aspect of the evaluation presented provides a stronger test of parameterisation performance compared with analysis of purely synthetic signals (e.g., in Strik et al., 1993; Strik, 1998). Although this dataset may on first inspection appear to be rather limited, the fact that each glottal pulse was manually analysed (both for inverse filtering and parameterisation) means that a considerable amount of effort was applied in order to have a reliable test set. To the best of the author's knowledge a dataset of this type and size has not been used previously for evaluating glottal source analysis techniques. Furthermore, these results were supported by analysis of larger datasets with less carefully controlled reference values.

It is rather difficult to determine the impact of the size of the errors observed for the various parameters. To assess this one would need to carry out analysis with specific applications, e.g., voice quality discrimination, parametric speech synthesis (such analysis is carried out in Chapter 6). Nevertheless, the qualitative evaluation provided some preliminary evidence that for the linguistic application of analysing glottal source parameters in various placements of focus, the proposed parameterisation technique was indeed suitable.

5.7 Conclusion

This chapter presents a new method (DyProg-LF) for automatic differentiated glottal source parameterisation which attempts to simulate some of the optimisation strategies used when the parameterisation is carried out manually and is guided by both time and frequency domain information. Results show that glottal source parameter data produced by this method were more similar to the reference data than those of the standard time-domain based method on a limited dataset of carefully controlled data. Furthermore, analysis of larger corpora supported the usefulness of the new method as it was also shown to estimate parameter values closer to the reference values than both the time-domain method and a phase minimisation based method. As a next step, it is intended to

apply a similar approach in the development of more robust automatic inverse filtering.

5.8 Applications

The proposed glottal source parameterisation method described in this chapter has strong potential for alleviating the workload of manual parameterisation approaches while still maintaining high precision. Consequently this method has been incorporated into the software (see Figure 5.1) used at the Phonetics and Speech Laboratory for glottal source analysis. The method may also be useful for including glottal source characterisation in statistical parametric speech synthesis and voice modification methods. Future research will involve exploiting this parameterisation method, in combination with the SE-VQ GCI detection method, for these very purposes.

Relevant publications

- Kane, J., Gobl, C., (2013) Automating manual user strategies for precise voice source analysis, *Speech Communication* 55(3), pp.397-414.

- Kane, J., Yanushevskaya, I., Ní Chasaide, A., Gobl, C., (2012) Exploiting time and frequency domain measures for precise voice source parameterisation, *Proceedings of Speech Prosody, Shanghai, China*, 143-146.

- Yanushevskaya, I., Gobl, C., Kane, J., Ní Chasaide, A., (2010) An exploration of voice source correlates of focus, *Proceedings of Interspeech, Makuhari, Japan*, 462-465.

Chapter 6

Evaluation of automatic glottal source analysis methods

Summary

This chapter documents a comprehensive evaluation carried out on automatic glottal inverse filtering and glottal source parameterisation methods. The experiments consist of analysis of a wide variety of synthetic vowels, of the ability of derived parameters to differentiate breathy to tense voice and a subjective evaluation of the perceptual quality of resynthesised utterances. One striking finding is that glottal model-based parameters compared favourably to parameters measured directly from the glottal source signal, in terms of separation of breathy to tense voice. Also, certain combinations of inverse filtering and parameterisation methods were more robust than others. For instance, closed-phase inverse filtering was shown to be effective for obtaining time domain based parameters, but considerably less so for the frequency domain parameter. Further results suggest that a recently proposed algorithm for fitting a glottal source model, when used for resynthesis, provided a more comparable quality to the original utterance than a standard method.

6.1 Introduction

The production of voiced speech can be considered as: the sound source created by the vibration of the vocal folds (glottal source) inputted through the resonance structure of the vocal tract and radiated at the lips. Most acoustic descriptions typically used in speech processing involve characterisation of mainly the vocal tract contribution to the speech signal. However, there is increasing evidence that development of independent feature sets for both the vocal tract and the glottal source components can yield a more comprehensive description of the speech signal.

Recent developments in speech synthesis (Cabral et al., 2011b; Degottex et al., 2012), voice quality modification (Vincent et al., 2005; Degottex et al., 2011a; O' Cinnéide et al., 2011), voice pathology detection (Drugman et al., 2009b) and analysis of emotion in speech (Lugger and Yang, 2007; Yanushevskaya et al., 2009; Lliev et al., 2010; Tahon et al., 2012) have served to highlight the potential of features related to the glottal source.

However, approaches for analysing the glottal source are at times believed to lack robustness in certain cases. For instance, higher pitch voices are known to be problematic for inverse filtering (Walker and Murphy, 2007) and particularly when combined with a low first formant frequency. This can result in strong source-filter interaction effects (Lin, 1990, 1987) which seriously affect the linear model of speech exploited in inverse filtering. Furthermore, precise glottal source analysis is often said to require the use of high-quality equipment to capture speech recorded in anechoic or studio settings (Walker and Murphy, 2007). Despite these claims, some studies have found that glottal source parameters derived from speech recorded in less than ideal recording conditions to contribute positively to certain analyses (Campbell and Mokhtari, 2003; Scherer et al., 2012b; Székely et al., 2012b).

It follows that the purpose of this chapter is to investigate performance of both inverse filtering and parameterisation steps, typically used in glottal source analysis. The evaluation of glottal source analysis methods is known to be problematic as it is not possible to obtain 'true' reference values. To deal with this, the current study presents a range of evaluation procedures in order to provide a more thorough impression of the performance of the various methods. Some similar work was recently carried out in Drugman et al. (2011) and the current study builds on this by incorporating model fitting methods and a subjective evaluation, involving copy-synthesis of recorded utterances.

6.2 State-of-the-art

A description of the state-of-the-art in terms of automatic glottal inverse filtering and glottal source parameterisation methods was previously given in Chapter 2, Section 2.3. For a more detailed review the reader can refer to Alku (2011) or Walker and Murphy (2007).

For the evaluation in the present chapter the glottal inverse filtering methods are evaluated: a closed-phase inverse filtering method (CPIF), iterative and adaptive inverse filtering (IAIF, Alku, 1992) and mixed-phase decomposition based on the complex-cepstrum (CCEPS, Drugman et al., 2009a). See Chapter 2, Section 2.3.1 for a description of these methods. Note that for these methods glottal closure instants (GCIs) are detected using the SE-VQ algorithm (Chapter 4). For the CPIF method GOIs are detected using the algorithm described in Drugman et al. (2012c).

The glottal source parameterisation methods are divided into two groups: direct measures and model fitting. The direct measures used in the current study are: the normalised amplitude quotient (NAQ, Alku et al., 2002), the quasi-open quotient (QOQ, Hacki, 1989) and the difference between the first two harmonics of the narrowband glottal source derivative spectrum (H1-H2). These three parameters were chosen as they were shown to be particularly effective at discriminating breathy to tense voice in a previous study (Airas and Alku, 2007) and are described in Chapter 2, Section 2.3.3.

Three algorithms are included which involve fitting LF model pulses to the glottal source derivative. A standard time domain method is used (Strik-LF, Strik, 1998), an amplitude based method (Amp-LF; see Chapter 2, Section 2.3.3) which exploits the equations developed in (Gobl and Ní Chasaide, 2003a) and an algorithm based on dynamic programming (DyProg-LF) described in Chapter 5.

One further algorithm is used in the evaluation which provides an estimate of the *Rd* parameter of the LF model by minimising a phase-based error criterion (Degott-LF, Degottex et al., 2011b). Again this method was previously described in Chapter 2, Section 2.3.3.

6.3 Experimental setup

The evaluation of glottal inverse filtering and glottal source parameterisation is problematic as neither source nor filter are directly observable and, hence, objective reference values cannot be obtained. Different evaluation methods have been proposed in the literature but each have their own shortcomings. The approach in this chapter is to carry

out a range of objective and subjective evaluation experiments in order to provide a more comprehensive impression of performance.

6.3.1 Synthetic testing

A frequently used evaluation procedure (see e.g., Strik, 1998; Sturmel et al., 2007; Drugman et al., 2011) is to do analysis of synthetic vowel segments where there are known reference values. This has the advantage of allowing straightforward quantitative evaluation where specific modifications to both vocal tract and glottal source model settings can be investigated. The disadvantage, however, is that the stimuli will be a simplified version of real speech and will not contain some of the known difficulties for glottal source analysis (e.g., the presence of aspiration noise, source-filter interaction effects, etc.).

In this chapter, analysis was carried out on a large range of synthetic vowel segments with wide variation of glottal source and vocal tract filter model settings. This was done in a similar fashion to that in Drugman et al. (2011). The range of variations is summarised in Table 6.1. The LF model was used to generate the synthetic source signal and was varied using the parameters f_0, Ra, Rk and Rg. With each setting 10 LF pulses were concatenated to create the source signal. An all-pole vocal tract model was used to modulate the source signal. Eight vowel settings were used based on the analysis of Finnish vowels (see Section 6.3.2) said by a single male speaker. The vowels were manually analysed using the procedure described in Section 2.3.5 and the derived formant frequencies and bandwidths were used as the vocal tract filter settings. In total 198,720 synthetic signals were generated for analysis. A small proportion of these variations resulted in improper LF model configurations [1]. These signals were discarded from the evaluation.

Table 6.1: Summary of glottal source and vocal tract parameter variations used in the synthetic signal testing.

	VOICE SOURCE				**VOCAL TRACT FILTER**
	f_0 **(Hz)**	**Ra**	**Rk**	**Rg**	**Vowel**
Max	300	0.15	0.5	2.0	9 Vowels settings
Min	80	0.01	0.1	0.6	
Step	10	0.02	0.05	0.1	

In order to evaluate the performance of automatic inverse filtering the following three parameters were considered: NAQ, QOQ and H1-H2. These parameters were calculated

[1] This occurred when $Rk > 2Rg - 1$ or when $Ra > 1 - \frac{1+Rk}{2Rg}$

from the synthetic source signal, as reference values. Then for each synthetic vowel the three inverse filtering methods: CPIF, IAIF and CCEPS, were used to estimate the source signal. From the outputted signal of these three algorithms the three glottal source parameters were then calculated. Relative errors scores were then computed for each parameter and then were analysed as a function of f_0 values and first formant frequency (F1, derived from the all-pole settings).

6.3.2 Voice quality differentiation

One useful application of glottal source analysis is to automatically differentiate voice qualities. Furthermore, as NAQ, QOQ, and H1-H2 have been shown to be suitable for separating breathy to tense voice (see for example: Airas and Alku, 2007; Alku et al., 2002; Hacki, 1989; Hanson, 1997) it seems reasonable to assume that quality of inverse filtering can be somewhat evaluated on the basis of how well the extracted glottal source parameter differentiates the voice quality. Such an approach has been used in previous studies (Kane et al., 2010; Drugman et al., 2011) and has the advantage of allowing quantitative evaluation on natural speech. However, the particular glottal source estimate is not directly evaluated and the application of extracted parameter values to voice quality differentiation is one step removed from the glottal source analysis itself.

The speech data used in this stage of the evaluation contained both steady vowels and continuous speech. For the **vowel dataset** recordings from 6 female and 5 male speakers aged between 18 and 48 years were used (also used in a previous study by Airas and Alku, 2007). The speakers were originally asked to produce eight Finnish vowels /ɑ e i o u y æ ø/ using breathy, normal and pressed phonation types. Participants received training in producing the voice qualities before recording. While conducting the recording, speakers were asked to repeat the utterance with stronger emphasis on the voice quality when it was necessary. Each utterance was repeated three times resulting in 792 speech segments. Audio was captured using high quality recording equipment (a unidirectional Sennheiser electret microphone with a preamp, LD MPA10e Dual Channel Microphone Preamplifier).

In order to ensure there were three independent sets of voice qualities and to resolve the potential ambiguity of the 'normal' voice quality label [2] perceptual screening was carried out on the vowel data. Three annotators, all experienced in voice quality research and familiar with Laver's labelling scheme (Laver, 1980), were randomly presented with vowel utterances using a web interface. Annotators were asked to label utterances on a five point Likert scale (i.e. breathy, breathy/modal, modal, modal/tense and tense). The scale gave

[2] A person's 'normal' voice quality could, for example, be inherently breathy.

the option of choosing breathy/modal and modal/tense as well as the three individual labels. This allowed annotators to indicate their uncertainty over the voice quality label if elements of two voice qualities were perceived. Analysis of the ratings showed an inter-rater agreement of $\kappa = 0.526$ (which indicates moderate agreement). In order to have three sets of independent voice qualities some thresholds were set to determine whether each utterance was to be included for the analysis. If the mean rating for the given sample was more than 0.75 (in numerical scores) away from its original voice quality label then the utterance was excluded. Vowels were also excluded if the standard deviation of its rating was more than 1, as this demonstrated disagreement on the part of the participants on the labeling to be used. This resulted in 314 of the 792 total samples being excluded, with 478 included for analysis. The reduced set of vowels displayed a considerably increased inter-rater agreement of $\kappa = 0.717$.

For each included vowel segment automatic inverse filtering was carried out using CPIF, IAIF and CCEPS, and parameterised using NAQ, QOQ and H1-H2. Furthermore, *Rd* and OQ parameters were derived from the model fitting by the **Strik-LF**, **Amp-LF** and **DyProg-LF** methods. Only the IAIF inverse filtering was used for this. *Rd* was also derived using **Degott-LF**, which does not require prior inverse filtering. Median parameter values were recorded for each utterance and the distributions of these values were examined as a function of the voice quality labels (i.e., breathy, modal and tense). An explained variance metric was then derived as the squared Pearson's R coefficient by treating median parameter values as the dependent variable and voice quality label as the independent variable. A similar evaluation procedure was carried out in Airas and Alku (2007).

For the **sentence dataset** all-voiced spoken sentences from two separate databases were compiled. The use of all-voiced sentences allowed evaluation independent of the effects of using automatic voicing decision algorithms. Furthermore, as voicing transitions often display characteristics associated with laxer phonation this would affect the results. The first set of sentences came from the read-VQ recordings, described in Chapter 4. The speech data from 6 speakers (3 male and 3 female) was selected from the database. Only the sentences produced with breathy, modal or tense voice were used in the present chapter and of these only the five sentences (from each voice quality set) which were all-voiced were included. Note that the perceptual screening used in Chapter 4 also applies here, and any sentences not meeting the criteria were also excluded here. In total 30 breathy, 30 modal and 20 tense voice sentences were included for analysis.

breathy, modal, tense, harsh, falsetto, creaky or other (only breathy, modal and tense were used for the current study). They were also asked to state whether they believed the

Table 6.2: Summary of speech data used in the voice quality discrimination experiments. Note that this speech data is also used for the experiments in Chapter 7

Speech type	Speakers	Male	Female	Utterances
Vowels	11	5	6	478
Sentences	6	3	3	80
Sentences	3	3	0	90
Total	20	11	9	648

chosen voice quality was produced: for *some*, for *most* or *throughout* the utterance, and whether they were: *very* confident, *quite* confident, or *not* confident that the label they had chosen was suitable. For the present study I opted to use only those utterances which were consistently given the 'correct' voice quality label, and where both annotators were *very* confident and believed the voice quality was maintained *throughout* the utterance.

Further all-voiced sentences were included, again with the same recording conditions as for the sentences above (described in Chapter 4, Section 4.5.1). 10 all-voiced sentences produced by 3 male speakers, in breathy, modal and tense voice, were recorded and added to the sentence dataset. The three male speakers were all experienced in voice-related research and individual utterances were re-recorded in several iterations until the sentences were deemed to properly represent the stated voice quality mode for the entire utterance.

A summary of the speech data used in the voice quality discrimination experiments is given in Table 6.2.

Again NAQ, QOQ, H1-H2, *Rd* and OQ parameters were derived as per the vowel dataset. This time, median values were not used as a parameter contour is more likely to vary substantially in continuous speech. However, in order to have a balanced dataset it is also desirable to have a fixed number of datapoints per sentence. To address this parameter contours were derived using each of the methods. These contours were then resampled to 10 samples which was deemed sufficient in order to capture any variations that might exist in the parameter contour but still maintaining a constant number of datapoints.

6.3.3 Perceptual testing

A final subjective evaluation was carried out involving glottal source and vocal tract parameterisation, and subsequent resynthesis. By doing perceptual evaluation of resynthesised utterances it is possible to evaluate the glottal source modelling as a whole, rather

than focusing on individual parameters.

For this stage of the evaluation 10 all-voiced sentences produced by a single male speaker from the sentence dataset was used. This speaker was chosen as he displayed a clearly modal voice quality with no audible aspiration noise. This was important as the glottal source model can only be used for characterising the deterministic component of the glottal source. Although aspiration noise models have been proposed (see e.g., Gobl, 2006; Degottex et al., 2011a) the aim here was to focus on the timbre resulting from the deterministic part of the glottal source.

Glottal inverse filtering and vocal tract modelling were carried out only using the IAIF method. The reason for not using CPIF was that occasionally spurious GOI markings can lead to a severe degradation in the quality of the resynthesis. Also, as the decomposition using CCEPS does not allow separation of the entire glottal source (i.e. open phase and return phase) from the vocal tract component, it did not allow easily comparable glottal source modelling.

Estimated glottal source derivative signals were parameterised using the two model fitting methods: **Strik-LF** and **DyProg-LF**. Parameter contours for the Strik-LF method were smoothed using a five-point moving average filter to lessen the effect of sudden changes in the model shape. Resynthesised utterances were generated by first creating a synthetic glottal source derivative using the parameters derived for the particular method. Voice source pulses were constructed and aligned to the corresponding GCI. Then the synthesised speech was created by filtering GCI-centred two pulse length glottal source derivative model frames with the corresponding all-pole filter. Successive frames were combined using overlap-and-add. Note that the same vocal tract filtering was applied for both methods.

Perceptual evaluation was carried out in the form of a modified ABX style design using a web application. 14 participants were first presented with the original utterance as a reference. They then listened to two resynthesised utterances and had to choose which sounded most like the original. This resulted in 10 ABX sets which the participants had to rate. The utterance order and the order of the methods were both randomised for each individual participant.

6.4 Results

6.4.1 Synthetic testing

The results from the synthetic testing are shown in Figures 6.1 and 6.2. The NAQ parameter was shown to be rather insensitive to variations in f_0 (Figure 6.1 a). Below around 240 Hz the IAIF method produces the lowest relative error, however from after this point the three inverse filtering methods yield similar results. Although these results corroborate previous findings in Drugman et al. (2011) for the performance of NAQ on synthetic data.

Also in terms of F1, NAQ appears to be insensitive to its variation. Again IAIF appears to provide the lowest relative error scores, although there is a sudden increase for the vowel setting with an F1 of 344 Hz. This can be explained by the fact that this was a /u/ vowel setting with a very low second formant. IAIF may at times have treated this as a single formant resulting in incomplete formant cancellation which affected the NAQ calculation.

For QOQ, the closed-phase inverse filtering method (CPIF) provided the lowest relative error scores. This was particularly true for higher f_0 values, with both IAIF and CCEPS showing significant increases in relative error from around 200 Hz. There was a clear affect of certain vowel settings on IAIF and CCEPS, but they were not clearly as a result of F1. CPIF was shown not to be affected by the different vowel settings.

In the case of H1-H2, however, CPIF gave clearly the highest relative error values. It was apparent from the analysis that even though the extracted time domain waveform, using CPIF, was suitable for deriving time domain parameters, it was considerably less so for the frequency domain one. The CPIF method seemed unable to obtain the relative amplitude of the first few harmonics.

6.4.2 Voice quality differentiation

The results from the voice quality differentiation experiments are shown in Figure 6.3 - 6.8 and Tables 6.3 - 6.6.

For the vowel dataset NAQ gave the best differentiation of the three voice qualities when the differentiated glottal source was estimated using the IAIF method ($R^2 = 0.60$). This trend corroborates previous analysis with a similar version of the vowel dataset (Airas and Alku, 2007). In fact in the present results the performance was even better than was previously reported. NAQ derived from CPIF and CCEPS gave a lower level of performance ($R^2 = 0.09$ and 0.24, respectively). By considering Figure 6.3 it is clear that this is largely due to a weaker separation of breathy and modal voice. For CPIF this may be explained by the difficulty in deriving the glottal opening instant (GOI) used for

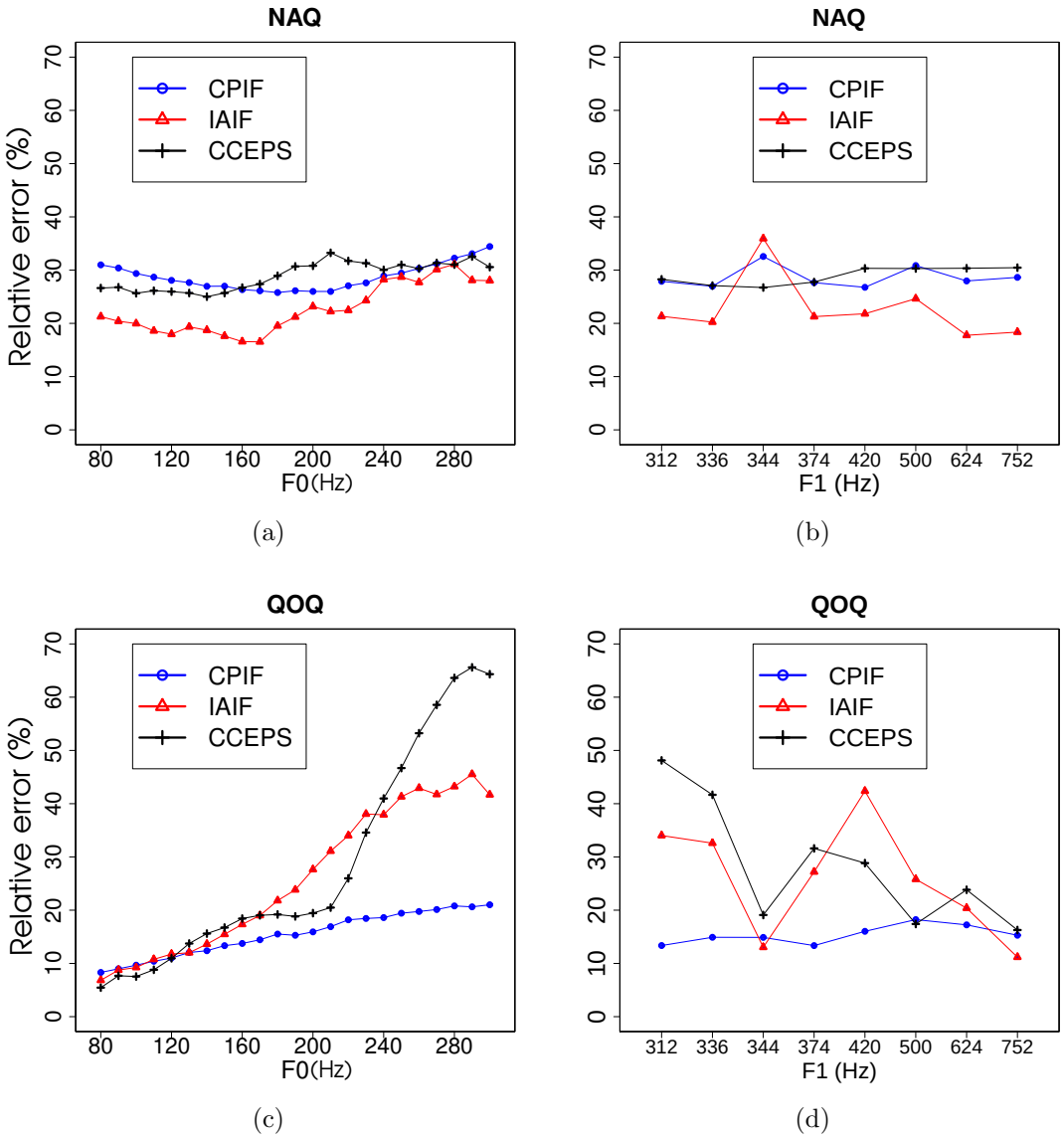
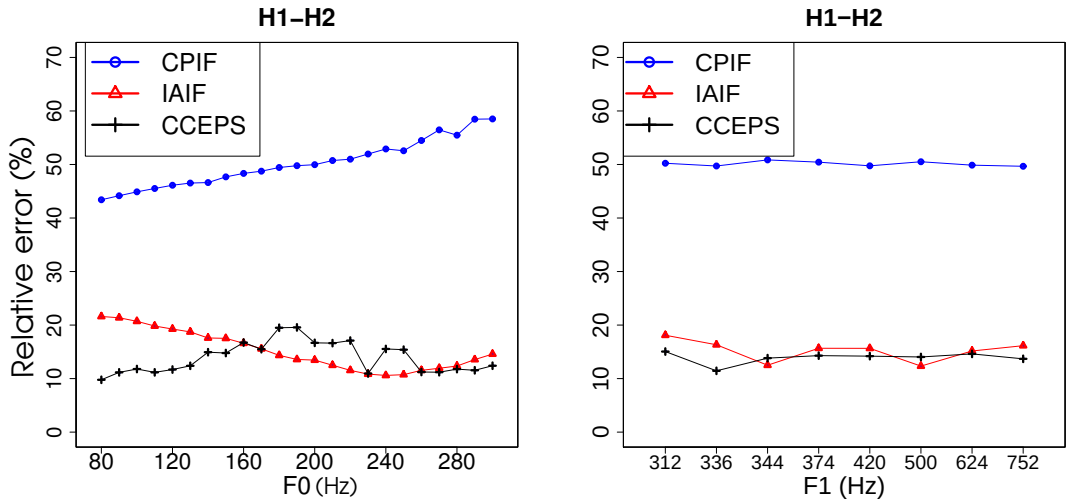


Figure 6.1: Mean relative error score for NAQ (top row) and QOQ (bottom row) as a function of f_0 (left column) and F1 (right column), for the three inverse filtering methods: CPIF (blue), IAIF (red) and CCEPS (black).

demarcation of the closed-phase. For CCEPS, which requires precise location and setting of the window function used, errors in GCI locations may have affected the analysis of breathy samples.

For H1-H2, the CCEPS decomposition was most suitable for allowing differentiation of breathy, modal and tense voice ($R^2 = 0.55$). Although IAIF provided weaker differentiation of modal and tense voice, both CPIF and IAIF, however, still resulted in useful H1-H2 values ($R^2 = 0.40$ and 0.30, respectively). For the parameter QOQ, all three



(a) (b)

Figure 6.2: Mean relative error scores for H1-H2 as a function of f_0 (left panel) and F1 (right panel), for the three inverse filtering methods: CPIF (blue), IAIF (red) and CCEPS (black).

decomposition methods provided similar voice quality differentiation performance.

Table 6.3: Explained variance (Pearson R^2) for each parameter and inverse filtering type combination. The glottal source parameter is treated as the dependent variable and voice quality label as the independent variable. Speech data comes from the **vowel dataset**.

	CPIF	**IAIF**	**CCEPS**
NAQ	0.09	0.60	0.24
H1H2	0.40	0.30	0.55
QOQ	0.44	0.42	0.42

Considering the differentiation of voice quality in the vowel dataset for the LF model fitting methods, the *Rd* parameter gave similar performance compared with NAQ. This was expected for Amp-LF ($R^2 = 0.65$), as the open-phase setting is derived following amplitude based measurements. However, Strik-LF and DyProg-LF gave *Rd* values ($R^2 = 0.62$ and 0.59, respectively) which performed considerably better than was previously reported (Airas and Alku, 2007). This difference is even more stark when considering OQ. It was also observed that the LF model fitting algorithm proposed in Chapter 5 (DyProg-LF) produced OQ values which differentiated voice quality ($R^2 = 0.56$) better than the direct measures QOQ and H1-H2. *Rd* derived using Degott-LF (see Figure 6.5) also provided

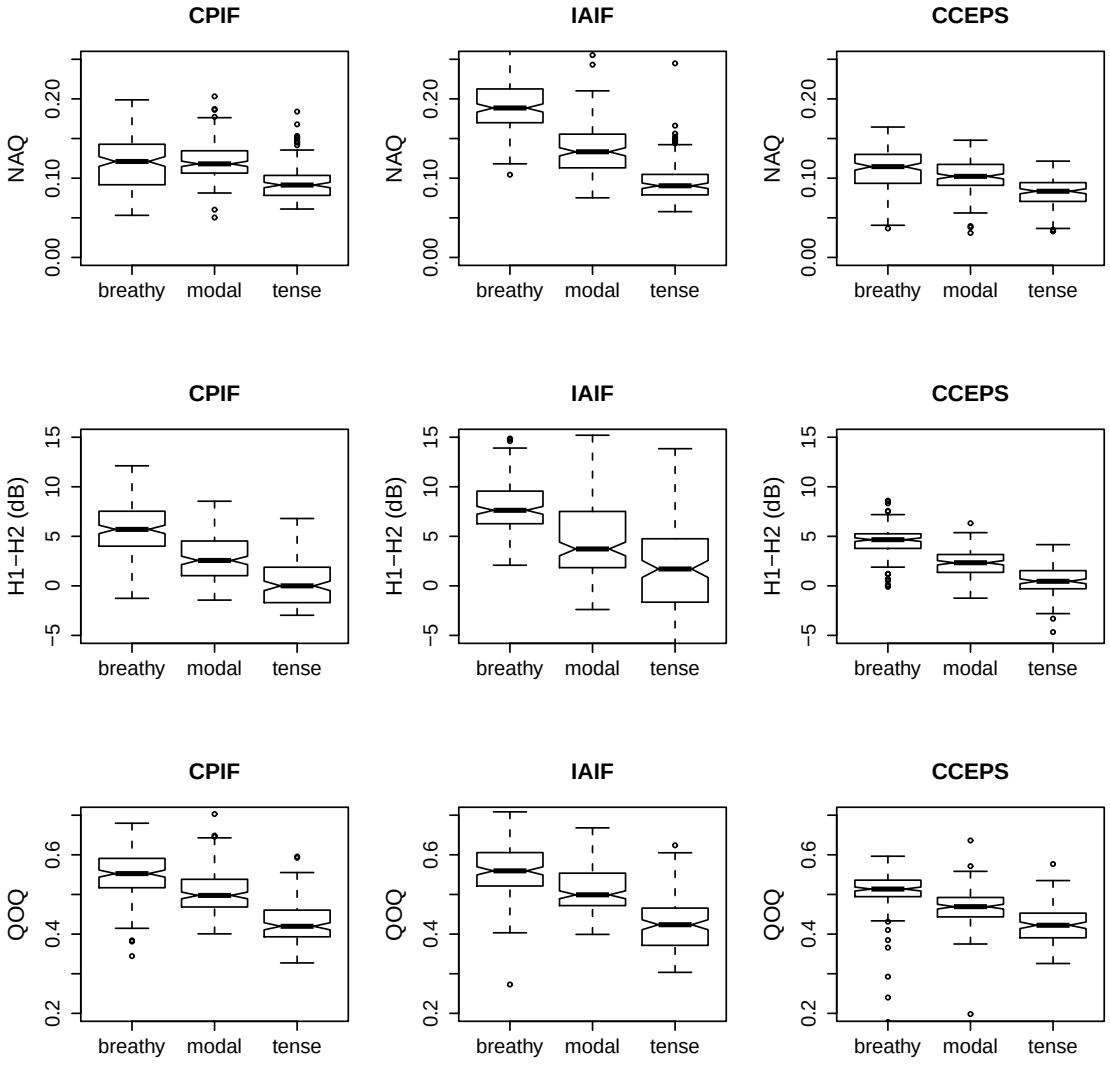


Figure 6.3: Distributions of NAQ (top row), H1-H2 (middle row) and QOQ (bottom row), for closed phase inverse filtering (left column), iterative adaptive inverse filtering (middle column) and complex-cepstrum based decomposition (right column), plotted as a function of voice quality. Speech data used is the **vowel dataset**.

very strong differentiation of breathy to tense voice ($R^2 = 0.62$)

As expected, overall differentiation of voice quality was reduced when moving to the combined sentence dataset. This is likely due to the difficulty in inverse filtering some parts of continuous speech (e.g., certain voiced consonants). However, similar trends were maintained with NAQ derived following IAIF giving the best performance ($R^2 = 0.22$). Once more CCEPS was the most suitable decomposition method for applying H1-H2 ($R^2 = 0.26$). For QOQ, a serious degradation in performance was observed for all decomposition

Table 6.4: Explained variance (Pearson R^2) for each LF model-based parameter (i.e *Rd* and *OQ*) and parameterisation type combination. The glottal source parameter is treated as the dependent variable and voice quality label as the independent variable. Speech data comes from the **vowel dataset**.

	Strik-LF	**Amp-LF**	**DyProg-LF**	**Degott-LF**
Rd	0.62	0.65	0.59	0.62
OQ	0.33	0.37	0.56	

methods.

Table 6.5: Explained variance (Pearson R^2) for each parameter and inverse filtering type combination. The glottal source parameter is treated as the dependent variable and voice quality label, from the **combined sentence datasets**, as the independent variable.

	CPIF	**IAIF**	**CCEPS**
NAQ	0.06	0.28	0.10
H1H2 (dB)	0.06	0.22	0.26
QOQ	0.09	0.20	0.05

The degradation in performance was notably less for the model fitting methods. This time the DyProg-LF method gave the best performing *Rd* values ($R^2 = 0.39$). This was also the case for OQ ($R^2 = 0.34$) and in fact both *Rd* and OQ derived from DyProg-LF provided considerably better voice quality differentiation than all the direct measure parameters. Another interesting observation was that the traditional OQ method consistently performed better than QOQ.

Table 6.6: Explained variance (Pearson R^2) for each LF model-based parameter (i.e *Rd* and *OQ*) and parameterisation type combination. The glottal source parameter is treated as the dependent variable and voice quality label, from the **combined sentence dataset**, as the independent variable.

	Strik-LF	**Amp-LF**	**DyProg-LF**	**Degott-LF**
Rd	0.21	0.26	0.39	0.28
OQ	0.24	0.20	0.34	

6.4.3 Perceptual testing

The results from the ABX-style perception test are shown in Figure 6.9. Participants signalled that synthesised utterances generated following the use of Dyprog-LF (83.6 % mean preference) were clearly more similar to the original utterance than when Strik-LF (16.43 % mean preference) was used. Note that the only difference between the synthesised utterances was the LF model shapes as derived using the two parameterisation methods. GCI positions, and hence f_0, and the vocal tract filtering were identical for the two methods. Listening to the resynthesised utterances it is clear that despite the moving average filter applied to the Strik-LF method, sudden changes in the glottal source settings degraded the quality somewhat. Furthermore, resynthesised utterances using Strik-LF were informally judged to be perceptually 'duller', suggesting inappropriate setting of the *Ra* parameter.

6.5 Discussion

Perhaps the most striking finding in this chapter was the strong performance of LF model based parameters at differentiating breathy to tense voice. Whereas the standard time domain LF model fitting algorithm (Strik-LF, Strik, 1998) gave comparable performance to that in Airas and Alku (2007), more recent algorithms for deriving LF model parameters (DyProg-LF, Kane et al., 2012 and Degott-LF, Degottex et al., 2011b) compared strongly with direct measure parameters.

This was particularly the case for continuous speech, where direct measure parameters suffered a serious degradation in performance. Specifically for DyProg-LF, both the *Rd* and OQ parameters still provided strong differentiation of the voice quality in continuous speech. The reason for the apparent robustness of the DyProg-LF method to continuous speech can be explained by the suitability of dynamic programming for maintaining sensible parameter contours even in *difficult* speech regions.

Although differentiation of voice quality does not directly measure the accuracy of derived parameter values, strong performance does suggest that the particular method is characterising salient glottal features.

Evidence from the testing on synthetic speech signals indicated that certain glottal inverse filtering methods were more suited to certain parameters. For instance, closed-phase inverse filtering (CPIF) was shown to be particularly suitable for deriving NAQ and QOQ, both time domain parameters. These parameters derived following CPIF were also rather insensitive to changes in f_0 and vocal tract filter setting. However, for the frequency domain parameter, H1-H2, the CPIF output was clearly less suitable. This finding may

corroborate those in Drugman et al. (2011) where CPIF was shown to produce higher levels of spectral distortion than the other inverse filtering methods.

However, the findings for IAIF conflict with those in Drugman et al. (2011), as in the present results IAIF had a similar performance to the other methods in terms of relative error on NAQ and QOQ, whereas in Drugman et al. (2011) it was considerably worse. In fact IAIF displayed relatively stable performance across the experiments and was shown to be particularly useful in combination with NAQ for breathy-tense discrimination and accuracy on synthetic speech signals.

Finally, results from the perceptual experiment with resynthesised utterances showed evidence that the recent parameterisation method, DyProg-LF, provided better modelling of the glottal source derivative than the Strik-LF method. The improvements are likely to have been brought about both by smoother parameter contours as well as better modelling of the higher frequencies. For this experiment an all-pole model was used for the vocal tract component for both methods as this facilitated focusing the participants on the effect of the glottal source modelling.

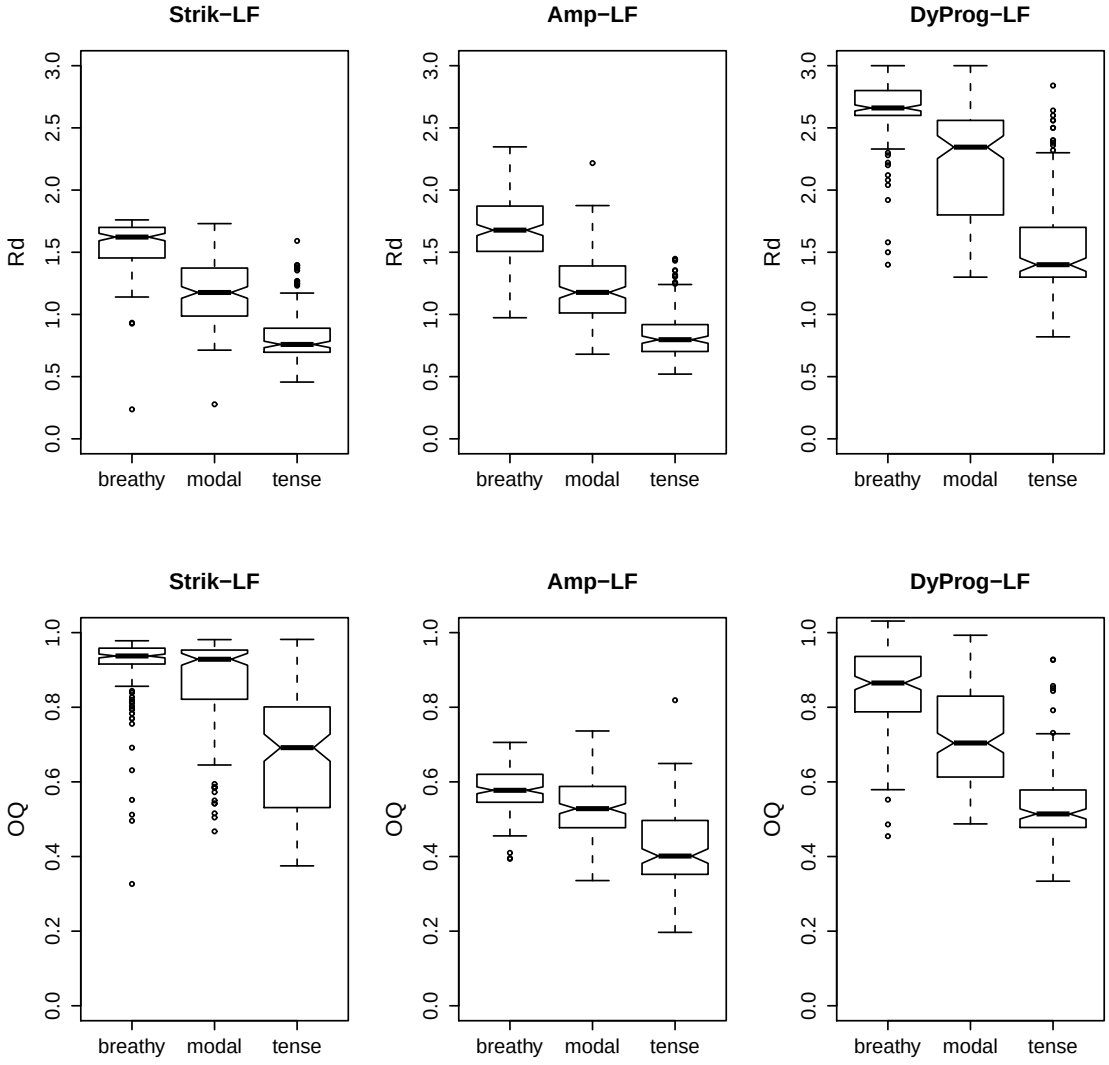


Figure 6.4: Distributions of *Rd* (top row) and OQ (bottom row) derived from the LF model fit, using the Strik-LF (left column), Amp-LF (middle column) and DyProg-LF (right column) algorithms and applied to the output of the IAIF method. Distributions are plotted as a function of voice quality label. Speech data used is the **vowel dataset**.

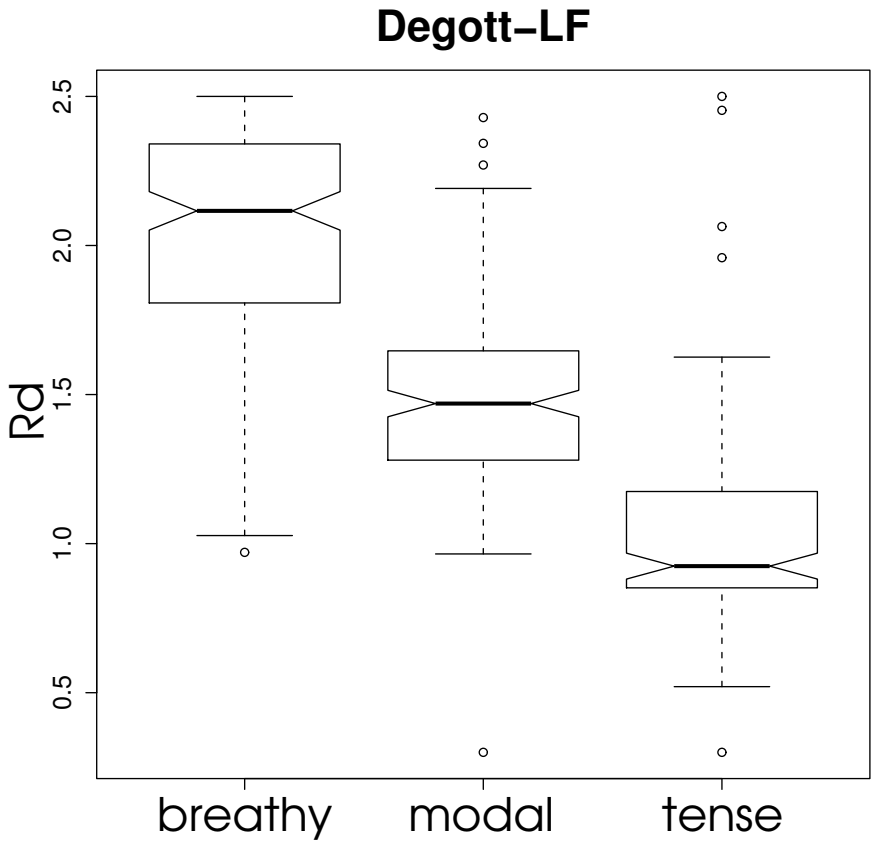


Figure 6.5: Distributions of *Rd* derived using the Degott-LF method. Distribution is plotted as a function of voice quality label. Speech data used is the **vowel dataset**.

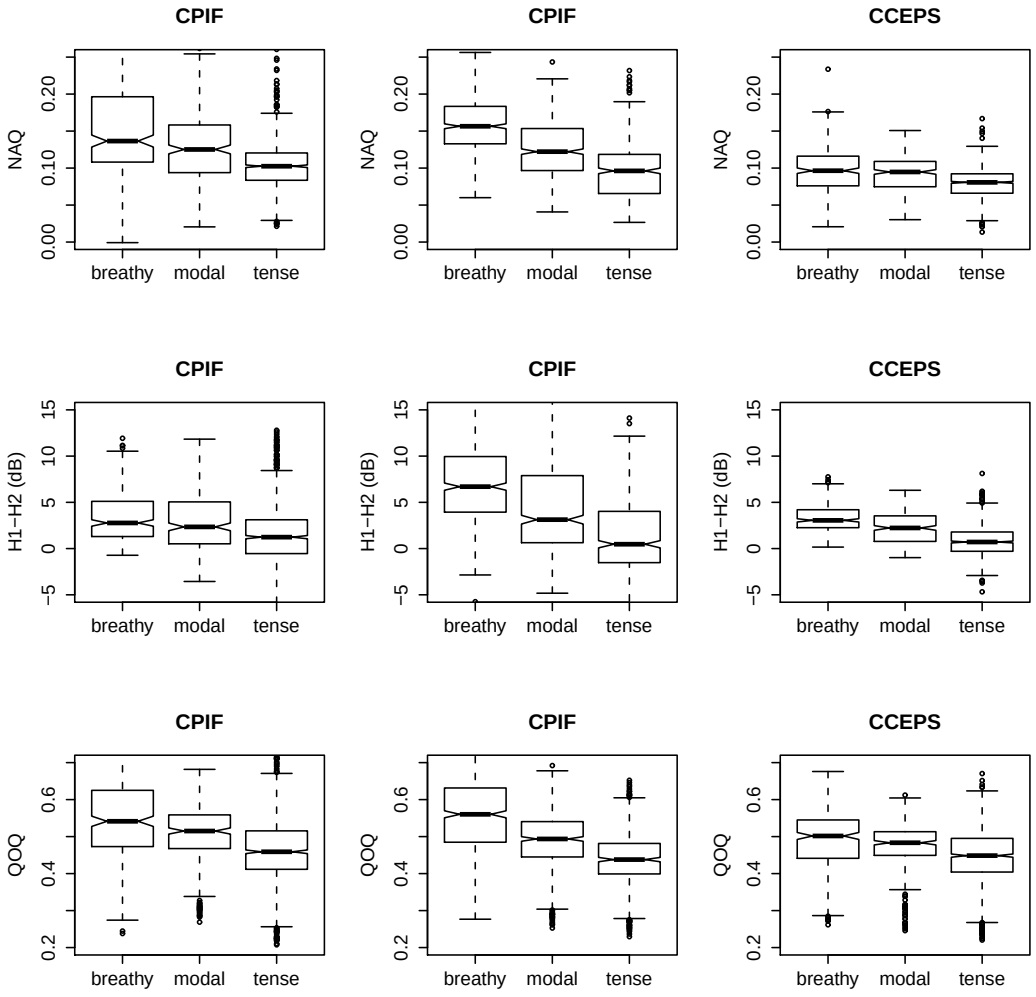


Figure 6.6: Distributions of NAQ (top row), H1-H2 (middle row) and QOQ (bottom row), for closed phase inverse filtering (left column), iterative adaptive inverse filtering (middle column) and complex-cepstrum based decomposition (right column), plotted as a function of voice quality. Speech data used is the **combined sentence dataset**.

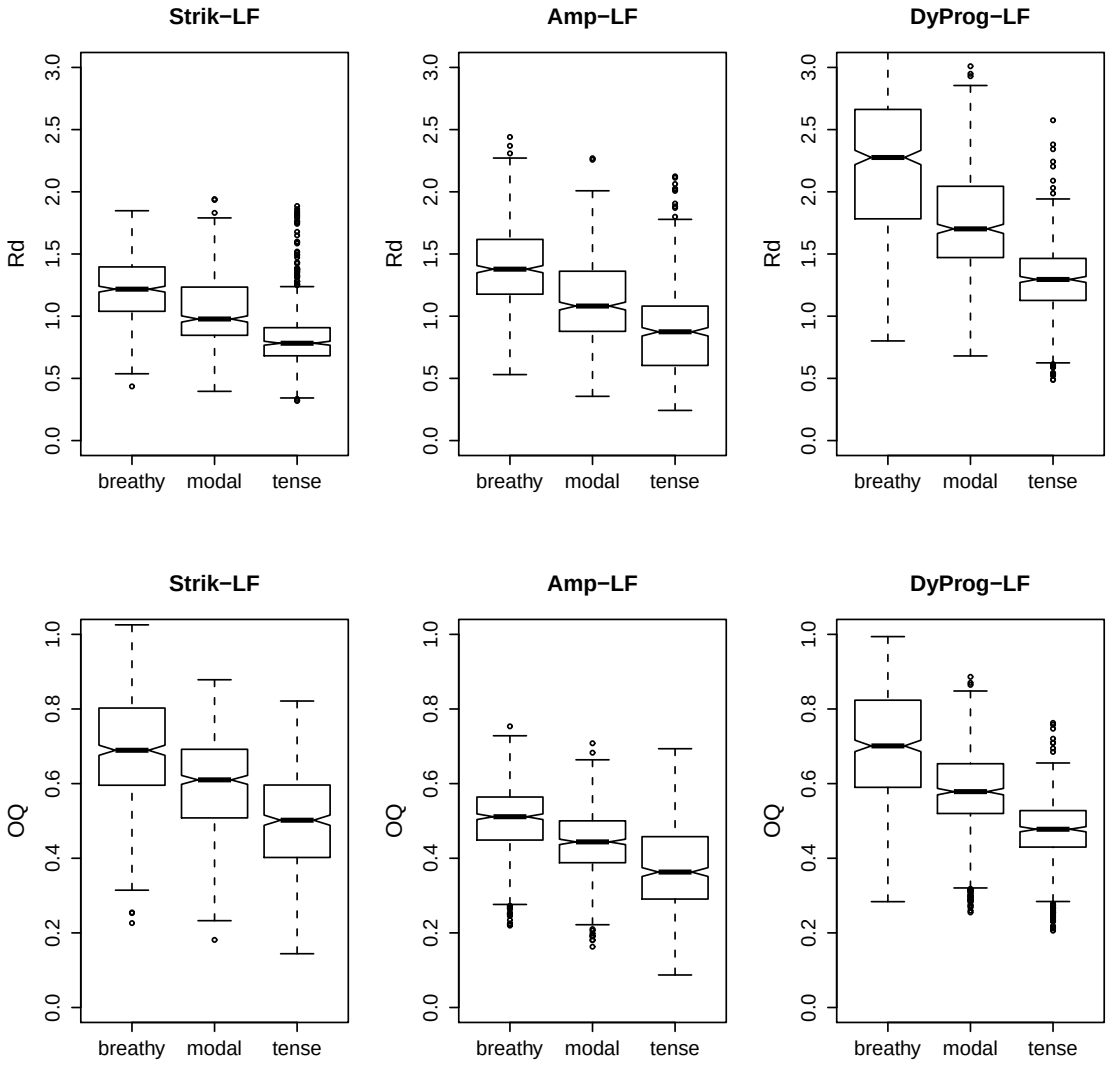


Figure 6.7: Distributions of *Rd* (top row) and OQ (bottom row) derived from the LF model fit, using the Strik-LF (left column), Amp-LF (middle column) and DyProg-LF (right column) algorithms and applied to the output of the IAIF method. Distributions are plotted as a function of voice quality label. Speech data used is the **combined sentence dataset**.

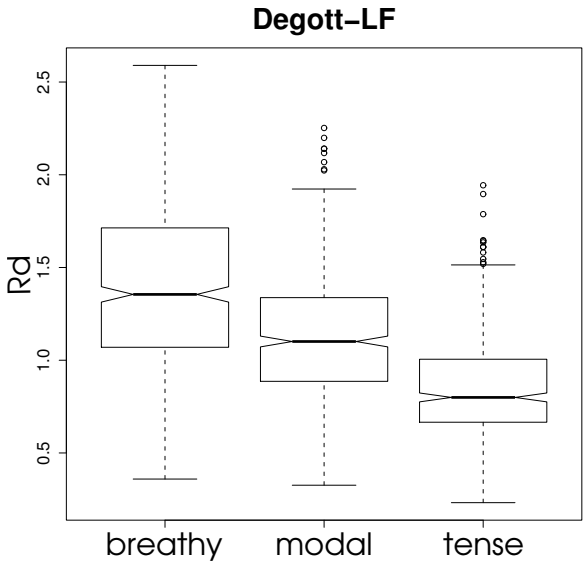


Figure 6.8: Distribution of *Rd* values derived using the Degott-LF method. Distribution is plotted as a function of voice quality label. Speech data used is the **combined sentence dataset**.

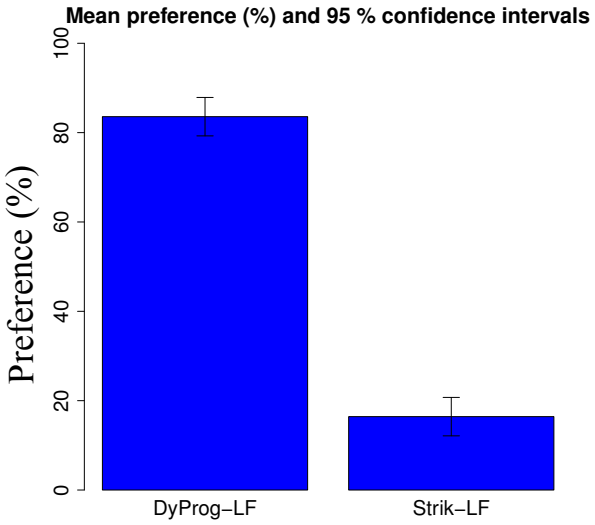


Figure 6.9: Mean percentage preference and 95 % confidence intervals for the two glottal source modelling methods used in the ABX style perception test.

For practical purposes it would be useful to remove the glottal source model from the speech signal (for instance using the glottal source separation method; Cabral et al., 2008), and characterise the resulting signal with a more sophisticated spectral envelope, e.g., STRAIGHT (Kawahara, 1997) or the True-Envelope (Villavicencio et al., 2006). A minimum-phase spectral envelope, like the True-Envelope and unlike the all-pole model, could characterise zeros occurring in the vocal tract filter, as occur in nasals. A similar approach to this has been used in parametric synthesis (Cabral et al., 2011b) and voice modification (Degottex et al., 2011a) using the LF model for generating the glottal source signal.

6.6 Conclusion

This chapter presents a general assessment of automatic glottal inverse filtering and glottal source parameterisation methods. To overcome the known difficulty of quantitative evaluation of glottal source analysis methods a range of different experiments were conducted which, in combination, provide a more comprehensive impression of the performance of the methods. Testing on synthetic signals revealed that different glottal inverse filtering methods were more suited to certain parameter estimation methods. The experiments on voice quality differentiation showed that more recent LF model fitting methods were as useful as direct measures for the vowel data and even more suited to the continuous speech data. Finally, the perceptual experiment revealed the recent parameterisation method DyProg-LF to be particularly suited modelling the glottal source. Resynthesised utterances were judged to be perceptually closer to the original utterance than a standard LF model fitting method.

Relevant publications

- Kane, J., Yanushevskaya, I., Ní Chasaide, A., Gobl, C., (2012) Exploiting time and frequency domain measures for precise voice source parameterisation, *Proceedings of Speech Prosody, Shanghai, China,* 143-146.

- Kane, J., Kane, M., Gobl, C., (2010) A spectral LF model based approach to voice source parameterisation, *Proceedings of Interspeech, Makuhari, Japan,* 2606-2609.

Part III

Coarse-grained analysis methods

Chapter 7

Wavelet maxima dispersion for breathy to tense voice discrimination

Summary

This chapter proposes a new parameter, the Maxima Dispersion Quotient (MDQ), for differentiating breathy to tense voice. Maxima derived following wavelet decomposition are often used for detecting edges in image processing, where locations of these maxima organise in the vicinity of the edge location. Similarly for tense voice, which typically displays sharp glottal closing characteristics, maxima following wavelet analysis are organised in the vicinity of the glottal closure instant (GCI). Contrastingly, as the phonation type tends away from tense voice towards a breathier phonation it is observed that the maxima become increasingly dispersed. The MDQ parameter is designed to quantify the extent of this dispersion and is shown to compare favourably to existing voice quality parameters, particularly for the analysis of continuous speech. Also, classification experiments revealed a significant improvement in the detection of the voice qualities when MDQ was included as an input to the classifier. Finally, MDQ is shown to be robust to additive noise down to a Signal-to-Noise Ratio of 10 dB.

7.1 Introduction

Voice quality can be considered as the timbre or auditory colouring of a person's speech (Laver, 1980). Breathy voice and tense voice, which are often considered to be opposite ends of a voice quality continuum (Gobl and Ní Chasaide, 1992), are perhaps the most studied aspects of voice quality for non-pathological speech. Breathy and tense voice, along with other aspects of a speaker's voice quality, are important features of paralinguistic signaling in speech and can provide the listener with information pertaining to the speakers affective state (Campbell and Mokhtari, 2003). For an illustration of this see Chapter 3, Section 3.1.1.

In terms of speech processing, the importance of robust detection of breathy and tense voice has been highlighted in Chapter 3, Section 3.2.

Many parameters have been proposed in the literature for discriminating breathy to tense voice and these have been discussed previously in this thesis (see Chapter 2 and Chapter 6). Further methods have been described for detecting breathiness from the amount of aspiration noise present in the signal (Ishi et al., 2011; Ito, 2004). These methods follow observations in Klatt and Klatt (1990), where the authors observed the third formant region to be considerably noisier in breathy voice than modal voice samples.

This chapter proposes a new parameter for differentiating breathy to tense voice by applying wavelet analysis for exploiting the different acoustic characteristics of these voice qualities. The parameter is compared against state-of-the-art parameters and is comprehensively evaluated through three sets of experiments which consider the ability to differentiate the three voice quality classes. The experiments are used to assess the ability to classify these voice qualities with a classifier using multiple input features and also to determine the robustness of the parameters to simulations of degraded conditions.

7.2 Proposed method

The proposed method for differentiating breathy to tense voice arises partly out of the observations made in Tuan and d'Alessandro (199) and d'Alessandro and Sturmel (2011). In these studies wavelet based zero-phase octave band filtering was carried out on the inputted speech signal. So-called Lines of Maximum Amplitude (LoMA) were subsequently derived which involved linking the maxima in the different outputted waveforms and this was then used for determining glottal closure instants (GCIs, Naylor et al., 2007). The LoMA have also shown to be useful for determining standard voice quality parameters such as; the open quotient, amplitude of voicing and spectral tilt (d'Alessandro and Sturmel,

2011).

In image processing, the maxima in signals outputted following wavelet-based filtering are often used for detecting edges (i.e. a contour in an image across which the brightness changes suddenly; Mallet, 1999). In fact wavelet analysis is in general suited to the detection of singularities in signals and indeed the characterisation of different types of singularities (Mallet and Zhong, 1992). The proposed method looks to exploit the different glottal closing characteristics of breathy, modal and tense voice by deriving a measurement following wavelet analysis.

A dyadic (i.e. based on powers of two) wavelet transform is carried out based on the method described in d'Alessandro and Sturmel (2011). A cosine-modulated Gaussian pulse, $m(t)$, similar to that applied in (d'Alessandro and Sturmel, 2011) was used as the so-called mother wavelet:

$$m(t) = -cos(2\pi f_n t) \cdot exp\left(-\frac{t^2}{2\tau^2}\right) \tag{7.1}$$

where the sampling frequency fs = 16 kHz, $f_n = \frac{fs}{2}$, $\tau = \frac{1}{2f_n}$ and t is time. As the main excitation in the glottal source pulse typically presents as a negative peak (assuming positive polarity of the signal), the minus sign ensures that the filtering will produce positive signal values corresponding to glottal closure. The wavelet transform, $y_i(t)$, of the input signal, $x(t)$, at the i^{th} scale s_i is calculated by:

$$y_i(t) = x(t) * m\left(\frac{t}{s_i}\right) \tag{7.2}$$

where $*$ denotes the convolution operator. The term scale refers to different scaled versions of the wavelet function, m, used in the convolution. Lower scales correspond to higher frequencies and vice versa. Here i = 0,1,2,....,6 was used which results in an octave band zero-phase filter bank, with filters having centre frequencies of: 8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz, 250 Hz and 125 Hz.

It is instructive to consider the output of this type of filtering on a negative Dirac impulse, see Figure 7.1. The maxima of the filter responses at the different frequencies are aligned as a result of the zero-phase filtering.

This observation is exploited in the proposed method by making the glottal excitation in tense voice analogous to a negative Dirac pulse. Furthermore, it is considered that as the phonation type tends away from tense voice and towards breathy voice that the glottal excitation becomes less and less like an impulse and in fact more like a sinusoid. Such an assumption is not unreasonable considering example tense and breathy settings of the

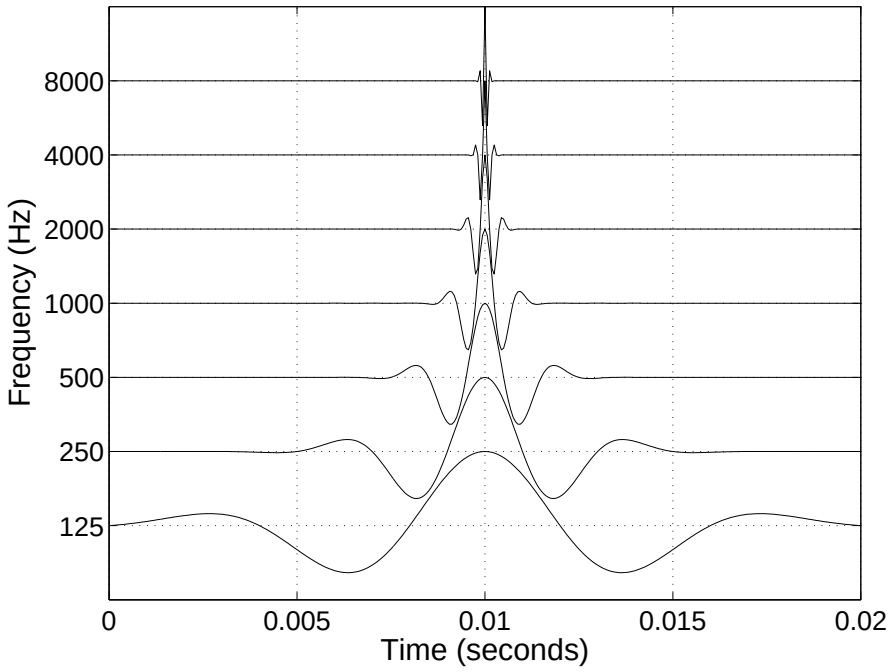


Figure 7.1: Output of wavelet decomposition for a negative Dirac pulse. Outputted signals are normalised in amplitude.

Liljencrants-Fant (LF, Fant et al., 1985a) differentiated glottal source model (Figure 7.2).

As was previously stated, in image processing when carrying out wavelet-based filtering the maxima of the outputted signals typically organise in the close vicinity of the location of the edge. This was also shown to be the case for a negative Dirac pulse (Figure 7.1). To illustrate what happens for breathy and tense voice, a spoken vowel sample is taken and the Linear Prediction (LP) residual is computed using autocorrelation LPC analysis and inverse filtering. This has the effect of removing the main oscillations resulting from formant resonance, while still maintaining the important phase properties of the glottal excitation. If one carries out the described wavelet filtering (Eq. 7.2) method on the breathy and tense residual signals and plot the maxima (i.e. any positive peaks) at the different scales one gets the output shown in Figure 7.3. For the breathy vowel (top panel) the maxima are extremely dispersed, whereas for the tense vowel (bottom panel) there is clear clustering of the maxima at regular intervals. In fact these clusters are located in the vicinity of the glottal closure instant (GCI). The proposed method involves quantifying the extent of this dispersion and a more formal description of the method now follows.

First, GCIs are detected using the SE-VQ algorithm (see Chapter 4 and Kane and Gobl, 2013). The LP-residual, $r_{LP}(n)$ (also used in SE-VQ), is decomposed using wavelet analysis, and is derived following autocorrelation Linear Predictive Coding (LPC) and

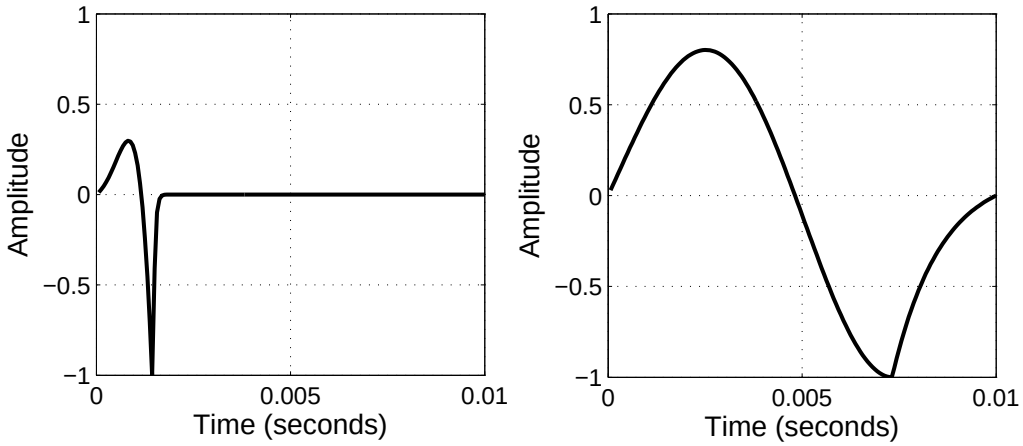


Figure 7.2: Example of an LF model pulse with *tense* (left panel) and *breathy* (right panel) parameter settings.

subsequent inverse filtering. A prediction order of $fs/1000 + 2$ is used which corresponds roughly to two coefficients to characterise each formant, for a typical male speaker, and two further coefficients to characterise the glottal source contribution. $r_{LP}(n)$ is then filtered using Eq. (7.2). Then for each GCI location a search interval, *int*, is defined as:

$$int = [GCI(k) - (T_0 \cdot c), \quad GCI(k) + (T_0 \cdot c)] \tag{7.3}$$

where k is GCI index and c is a constant which determines the size of the search region relative to the glottal period, T_0. Here c is set to 0.2 which allows maxima to be measured in the vicinity of the GCI while generally avoiding measuring maxima arising from neighbouring glottal pulses. The search interval is used to determine the location of maximum amplitude, m_i, in the vicinity of the GCI at each scale:

$$m_i = \arg\max_{y_i}\{y_i(int)\} \quad \text{for} \quad 0 \leq i \leq 6 \tag{7.4}$$

and from this the distance from these maxima locations, d_i, to the GCI can be measured:

$$d_i = |GCI(k) - m_i| \quad \text{for} \quad 0 \leq i \leq 6 \tag{7.5}$$

Finally, the maxima dispersion quotient, MDQ, is then calculated using:

$$MDQ(k) = \frac{\frac{1}{I}\sum_{i=0}^{I-1} d_i}{T_0(k)} \tag{7.6}$$

where I is the number of scales and $T_0(k)$ is the local glottal period duration.

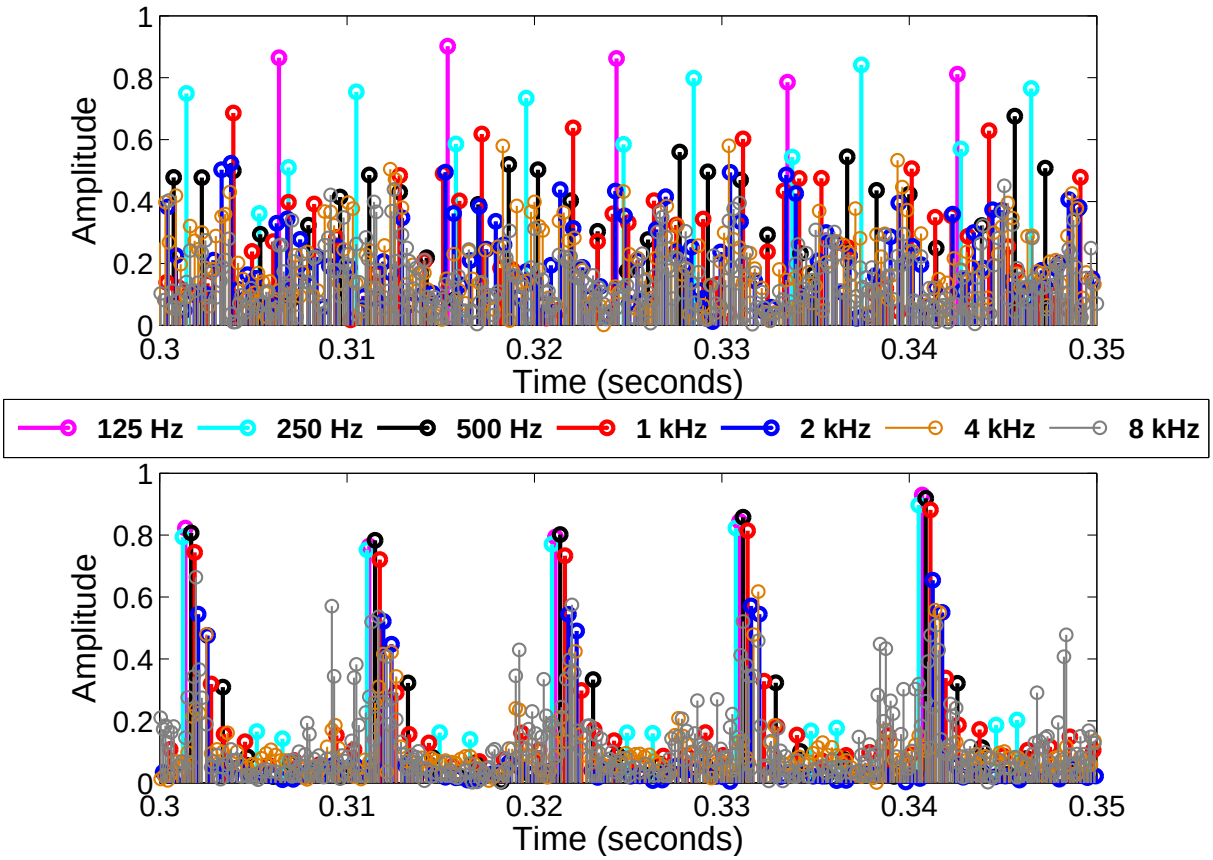


Figure 7.3: Illustration of dispersion of maxima measured following wavelet based decomposition of the LP-residual of a breathy vowel (top panel) and a tense vowel (bottom panel).

Note that this method for deriving MDQ is carried out on a glottal pulse-synchronous basis. For certain applications, however, one may wish to avoid GCI detection which may negatively affect the robustness of the measurement. In such cases MDQ could be calculated on a fixed frame-basis where the location of the maximum amplitude at the highest scale (lowest frequency) could be substituted for the GCI location. d_i could then be measured as distances from this point, and the denominator of Eq. (7.6) could be substituted with a $T_{0,mean}$ value. The author has observed this approach to produce broadly similar results to the standard MDQ calculation. However, it is the glottal pulse synchronous approach which is used throughout this chapter.

Figure 7.4 provides an illustration of the MDQ contour for an /a/ vowel produced by a male speaker beginning in a tense phonation and moving gradually to breathy voice. The MDQ contour moves steadily from a low value, around 0.04, to a higher value, around 0.09, in the breathy part. It is interesting to note how the MDQ contour increases as the

amplitude of the speech signal decreases. This is despite the calculation of MDQ being independent of variation solely in signal amplitude.

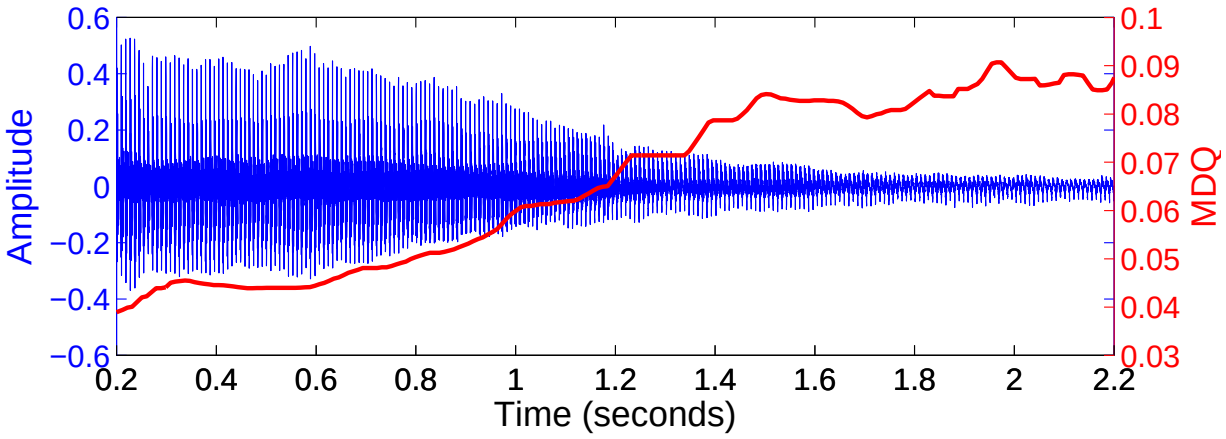


Figure 7.4: Speech waveform of an /a/ produced by a male speaker with the phonation type varying gradually from tense to breathy. Also shown is the extracted Maxima Dispersion Quotient (MDQ) contour.

7.2.1 Summary

The MDQ parameter measurement is summarised as follows:

1. Estimate GCIs using the SE-VQ algorithm (see Chapter 4 and Kane and Gobl, 2013)
2. Derive LP-residual signal, r_{LP}, using LPC analysis of order $fs/1000 + 2$
3. Carry out wavelet decomposition of r_{LP} using Eq. (7.2)
4. Define search interval relative to a given GCI location, Eq. (7.3)
5. Find locations of maxima, m_i, at the different scales in this search region (Eq. 7.4)
6. Measure the distance, d_i, of the maxima locations, m_i, to the given GCI location (Eq. 7.5)
7. Compute the average of these distances normalised to the glottal period, Eq. (7.6)

7.3 Experimental setup

This section describes the experimental setup, involving three sets of experiments, used to evaluate the proposed parameter, MDQ.

7.3.1 Speech data

Note that the speech data used in the current study was selected in order to have datasets with clearly discrete voice quality classes. Although larger corpora, such as that used in Schroeder and Grice (2003), may be useful for the objective of evaluating the ability of parameters at distinguishing similar voice quality classes to those considered here, the objective in this study was to ensure that there was as little overlap as possible between the voice quality labels.

The speech data used here are the same data that were used in Chapter 6, Section 6.3.2 and a summary of the data is shown in Table 6.2

7.3.2 Comparison parameters

As in Chapter 6, three parameters were used: the normalised amplitude quotient (NAQ, Alku et al., 2002), the quasi-open quotient (QOQ, Hacki, 1989) and the difference between the first two harmonics of the narrowband glottal source derivative spectrum (H1-H2), as comparison. Again these were chosen as they were shown to be particularly effective at discriminating breathy to tense voice in a previous study (Airas and Alku, 2007) and the description of them is given in Chapter 2, Section 2.3.3.

7.3.3 Experiments

A description of the three sets of experiments carried out in the current study is now given.

Differentiation of breathy to tense voice

The first set of experiments involved examining the ability of individual parameters to differentiate breathy to tense voice in the speech data. For the vowel dataset, each of the parameters were calculated and then the mean of each parameter was saved as the datapoint. For the dataset of sentences, however, parameter values are likely to vary considerably more and taking a single mean value is unsuitable. To handle this, the parameter contour was extracted for a given sentence and then resampled to 10 datapoints.

This allowed the variation in the sentence to be captured while still maintaining a balanced dataset.

The extent to which the individual parameters differentiate the three voice qualities is then demonstrated by plotting the distribution of each parameter as a function of voice quality and calculating Explained Variance. The metric Explained Variance is measured by treating parameter values as the dependent variable and voice quality label as the independent variable and calculating Pearon's R^2. This was done separately for the vowel and sentence datasets. This is the same experimental setup as was used in Chapter 6.

Classification experiments

Next the aim was to evaluate the extent to which the proposed parameter, MDQ, brings improvement to the classification of the three voice qualities. To do this Support Vector Machines (SVMs) were used as the classifier utilising a Radial Basis Function (RBF) kernel (Bishop, 2006). As there are three voice quality classes, a one-against-one multi-class architecture was opted for. 10-fold cross-validation experiments were conducted, where the dataset was randomly partitioned into 10 equal-sized sets. One fold was held out to be used solely for testing with the remainder of the dataset used for training. This was repeated for each of the 10 folds with classification error scores saved each time. These experiments were carried out for 5 different feature vectors:

1. MFCCs + f_0
2. MFCCs + f_0 + VQ(-MDQ)
3. MFCCs + f_0 + VQ
4. f_0 + VQ(-MDQ)
5. f_0 + VQ

where 13 mel-cepstral coefficients (MFCCs) were measured using 25 ms Hanning windowed frames, with a 5 ms shift, VQ is the full set of voice quality parameters (i.e. {MDQ, NAQ, QOQ, H1-H2}) and (-MDQ) indicates the VQ set with MDQ excluded. The corresponding voice quality class (i.e. breathy, modal or tense voice) was used as the target class. To examine the effect of the different feature vectors on the classification error, a one-way ANOVA was carried out with classification error treated as the dependent variable and feature vector type as the independent variable. Pairwise comparisons were computed using Tukey's Honestly Significant Difference (HSD) test.

Note that the primary concern here is not with the overall classification accuracy but rather the goal is to investigate whether the MDQ parameter can bring improvement to the accuracy.

Robustness of parameters to degraded conditions

The final set of experiments aimed at investigating the robustness of the four parameters to simulated degradations of the recorded speech signals. A similar cross-validation experiment to that in Section 7.3.3 was carried out. However, this time the analysis was conducted for each voice quality parameter separately (i.e. a one-dimensional feature vector each time). As previously, one of the 10 validation folds was held out and the classifier was trained on the remaining data. But for this analysis, the testing (i.e. on the one held out fold) was done on parameter values measured on the test set with noise added to the speech signals. The analysis was again repeated for each validation fold. Experiments were carried out adding white Gaussian noise and Babble noise (taken from Varga and Steeneken, 1993) to the signals at Signal-to-Noise Ratios (SNR) varying from 80 dB (almost 'clean' speech) to 0 dB (heavily degraded). Results of the analysis were determined separately for white noise and babble noise.

7.4 Results

7.4.1 Voice quality discrimination

The distributions of parameter values are plotted as a function of voice quality for the vowel dataset in Figure 7.5 and for the sentence dataset in Figure 7.6. Explained variance scores are also given for the two datasets in Table 7.1.

Table 7.1: Explained variance (Pearson R^2) for each parameter. The parameter is treated as the dependent variable and voice quality label, from the vowel and sentence datasets, as the independent variable.

Dataset	**MDQ**	**NAQ**	**QOQ**	**H1-H2**
Vowel	0.59	**0.60**	0.42	0.30
Sentence	**0.39**	0.28	0.20	0.22

For the vowel dataset, MDQ (R^2 = 0.59) and NAQ (R^2 = 0.60) clearly provided the best discrimination of the three voice qualities. For QOQ, breathy and modal voice

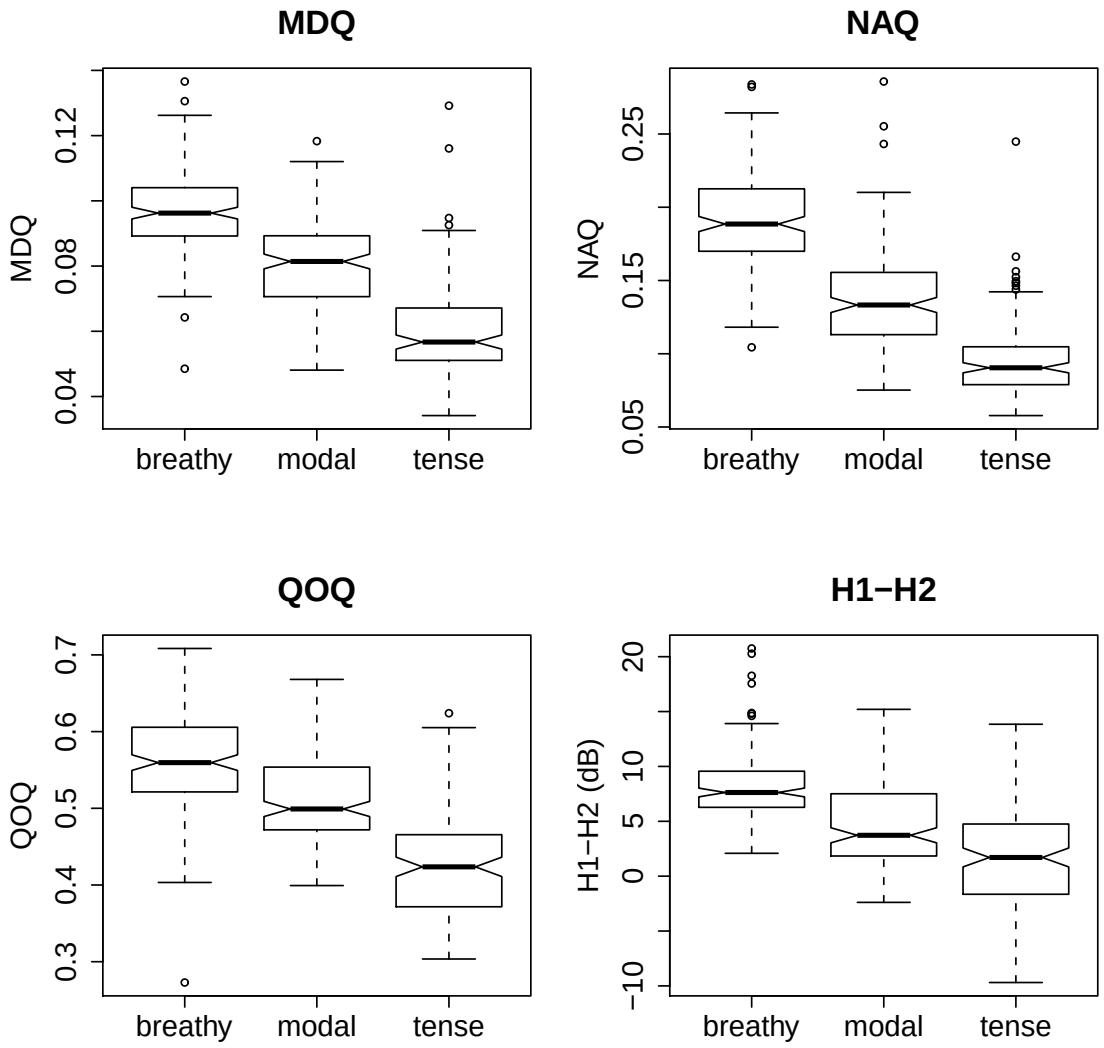


Figure 7.5: Distributions of MDQ (top left), NAQ (top right), QOQ (bottom left) and H1-H2 (bottom right) plotted as a function of voice quality for the **vowel** dataset.

were less clearly discriminated, while for H1-H2 the parameter failed to provide strong separation of modal and tense voice. The relative difference between the three standard parameters; NAQ, QOQ and H1-H2, in the current study corroborates previous findings in Airas and Alku (2007) which used the same vowel dataset. However, in the present study the three parameters produce clearly better discrimination than was reported in Airas and Alku (2007). This may be explained by the extra perceptual screening carried out on the dataset to ensure there were three separate voice quality classes (the reader can refer back to Chapter 6, Section 6.3.2 for details on this perceptual screening).

When considering the sentence dataset there is, not surprisingly, a general reduction

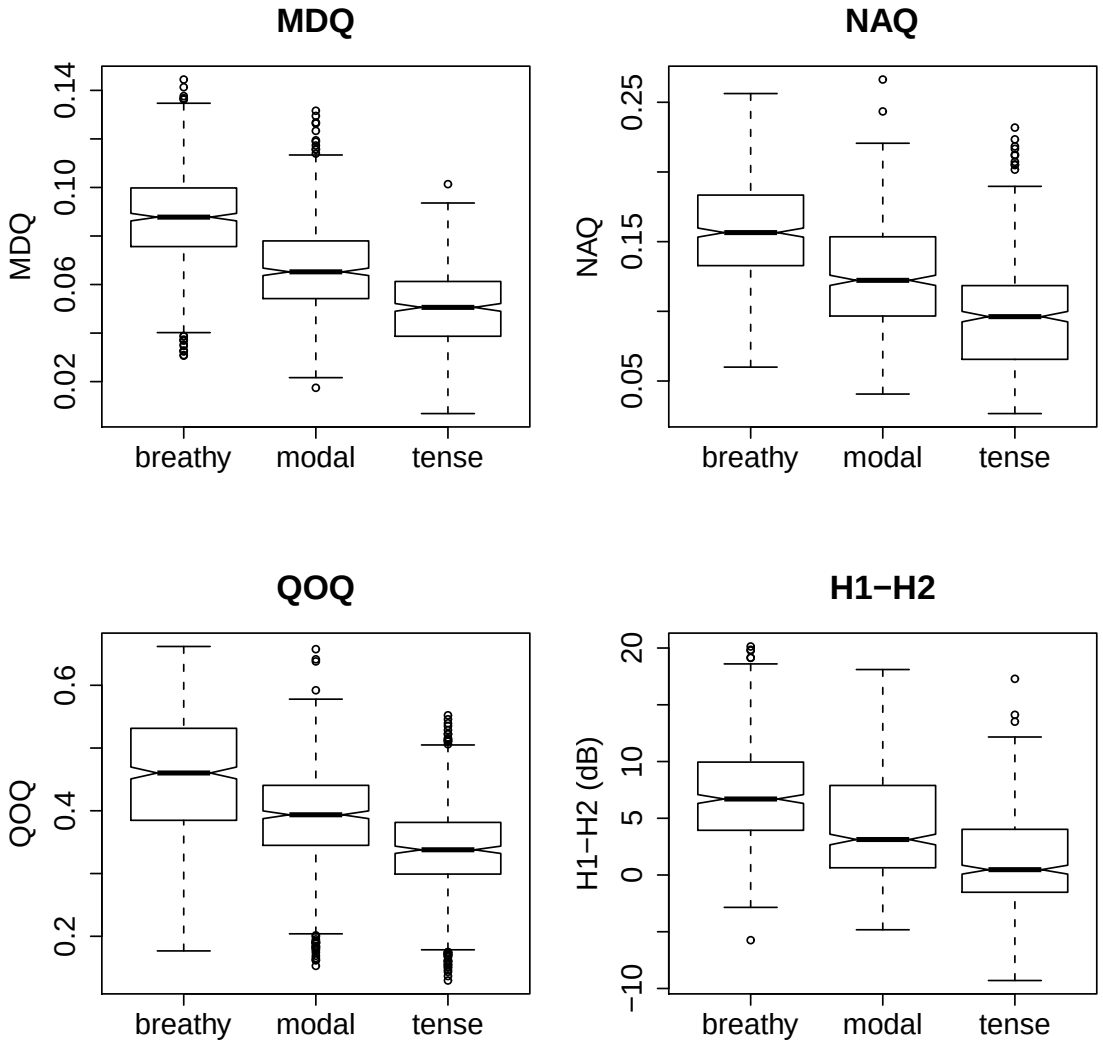


Figure 7.6: Distributions of MDQ (top left), NAQ (top right), QOQ (bottom left) and H1-H2 (bottom right) plotted as a function of voice quality for the **sentence** dataset.

in the discrimination of the three voice quality classes. Here MDQ clearly provides the best separation of the classes ($R^2 = 0.39$). NAQ ($R^2 = 0.28$) and QOQ ($R^2 = 0.20$) in particular show a dramatic degradation in performance, compared to the vowel dataset. Although the perceived voice quality was very stable across these utterances, it is known that the glottal source will nevertheless vary considerably in running speech (Gobl, 1988; Yanushevskaya et al., 2010). This would naturally affect the stability of the parameter values even if a fairly constant voice quality is perceived. Furthermore, running speech, compared to steady vowels, is more likely to cause problems for glottal inverse filtering. As the three comparison parameters are calculated following glottal inverse filtering this

may have caused the reduction in performance for running speech.

7.4.2 Classification experiments

The results from the 10-fold cross-validation classification experiment are shown in terms of mean and standard deviation of the classification error in Table 7.2 with a summary of pairwise comparisons following Tukey's HSD test shown in Table 7.3. The one-way ANOVA revealed a significant effect of feature vector type on the classification error [$F_{(4,45)} = 50.017$, $p < 0.001$]. In Table 7.2 it can be seen that including all parameters (i.e., MFCCs, f_0 and all VQ parameters, including MDQ) gave the lowest average classification error (20.2 %). Note that this was significantly lower ($p < 0.05$) than for the same feature vector just excluding MDQ. The confusion matrices shown in Table 7.4 show that this improvement involved an even reduction in classification error for the three voice quality classes. Furthermore, when excluding MFCCs, the feature vector including MDQ gave a significantly lower classification error ($p < 0.001$) compare to when it was removed.

Table 7.2: Mean and standard deviation of classification error scores (in %) following 10-fold cross validation experiments with different input feature vectors. Best performance is highlighted in bold and * denotes significant difference ($p < 0.05$) between the lowest error and the next lowest, following ANOVA and subsequent Tukey's HSD posthoc testing.

	10-fold cross validation	
Input features	Mean Error (%)	Standard deviation (%)
MFCCs + f_0	30.46	2.90
MFCCs + f_0 + VQ(-MDQ)	25.20	3.96
MFCCs + f_0 + VQ	**20.20***	2.63
f_0 + VQ(-MDQ)	40.76	3.37
f_0 + VQ	31.25	4.06

Another interesting finding was that the inclusion of voice quality parameters brought a significant improvement ($p < 0.001$) to the classification error when used in combination with MFCCs and f_0, compared to just MFCCs and f_0 alone.

These are positive findings for the proposed parameter, indicating that further information to do with breathy and tense voice is provided with MDQ, which is not captured within the existing parameters. The overall classification accuracy may be improved by including other voice quality parameters from the literature as inputs to the classifier. Furthermore, a recent study demonstrated that by using a large voice quality feature set and exploiting the disagreement on the part of the voice quality annotation, considerable

Table 7.3: Summary of pairwise comparisons from Tukey's HSD test, following a one-way ANOVA with classification error from 10-fold cross validation experiment as the dependent variable and input feature vector type as the independent variable. Significant pairwise differences are shown with * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$). Note that f_0 was also included in each of the feature vectors, however it has been omitted from the labels here for conciseness.

	MFCCs, VQ(-MDQ)	MFCCs, VQ	VQ(-MDQ)	VQ
MFCCs	*	***	***	0.98
MFCCs, VQ(-MDQ)		*	***	**
MFCCs, VQ			***	***
VQ(-MDQ)				***

Table 7.4: Confusion matrices (in %) following 10-fold cross validation experiments. The left matrix shows the results when the input feature vector consisted of all parameters with the exception of MDQ, whereas the right matrix shows results when MDQ is also included. Best voice quality classification on the diagonal of the confusion matrices are highlighted in bold.

	All features - MDQ			**All features + MDQ**		
	Breathy	Modal	Tense	Breathy	Modal	Tense
Breathy	77.9	20.0	2.1	**83.1**	16.0	0.9
Modal	17.3	68.3	14.4	13.1	**74.8**	12.1
Tense	3.8	17.4	78.8	2.4	15.8	**81.8**

improvements can be made in terms of accurate detection of these voice qualities (Scherer et al., 2012a).

7.4.3 Robustness testing

The effect of adding white and babble noise to the speech samples in the test set on the classification accuracy of the 10-fold cross validation experiments with single-dimensional input feature vector is shown in Figure 7.7.

The MDQ parameter achieved the highest classification accuracy in both white and babble noise conditions down to a Signal-to-Noise Ratio (SNR) of 20 dB. At 10 dB SNR the accuracy drops moderately for white (51 %) and babble noise (54 %). For the most severe noise condition, i.e. 0 dB SNR, the accuracy drops severely down close to chance levels (around 40 %). NAQ displays a similar trend, but with lower classification accuracy

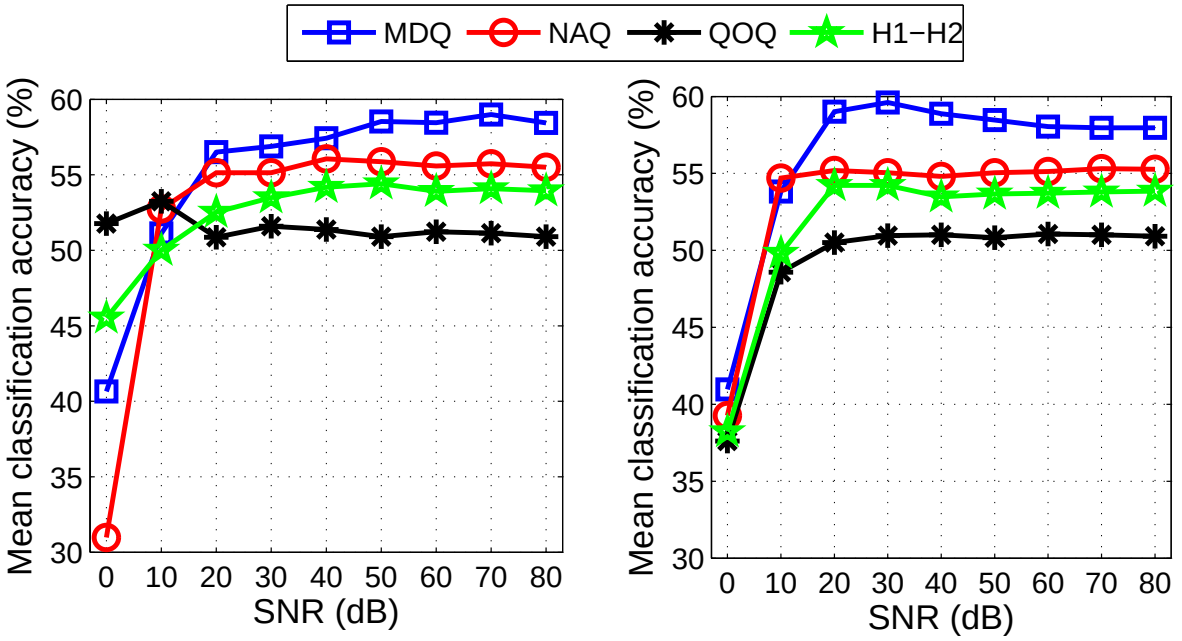


Figure 7.7: Effect of adding white noise (left panel) and babble noise (right panel) at varying Signal-to-Noise Ratios (SNR) on mean classification accuracy (%) in the 10-fold cross validation experiment.

than MDQ above 10 dB SNR. QOQ and H1-H2 both produced generally lower accuracy than MDQ and NAQ, except for 0 and 10 dB SNR in the white noise condition. For babble noise, however, the accuracy is severely degraded at 0 dB SNR. This is likely due to the more pronounced low frequency characteristic of babble noise, compared to white noise, which affected the calculation of these parameters.

7.5 Discussion and conclusion

This study presents a new parameter, the maxima dispersion quotient (MDQ), for discriminating breathy to tense voice. A comprehensive evaluation revealed that the new parameter provides comparable differentiation of the voice quality classes for the vowel dataset and clearly better differentiation for the dataset of running speech than the best performing comparison parameter (i.e. NAQ). The three comparison parameters were all calculated from the glottal source signal estimated by automatic inverse filtering (IAIF, Alku, 1992). However, automatic glottal inverse filtering of running speech can be problematic, particularly for certain voiced consonants. This may have affected the measurement of these parameters and reduced their ability to discriminate the voice qualities. MDQ, on the other hand, does not require glottal inverse filtering, but simply applies standard

LPC inverse filtering to remove the effect of oscillations emanating from both the glottal source contribution and from vocal tract resonance. This perhaps suggests that it is more suited to the automatic analysis of running speech.

Results from the classification experiment clearly demonstrated that MDQ provides further information for the discrimination of the three voice quality classes than is present in the three comparison parameters as well as the full set of mel-cepstral coefficients. While findings from the robustness testing suggested that MDQ can provide meaningful information for the discrimination of breathy and tense voice down as far as 10 dB SNR.

7.6 Applications

The new parameter, MDQ, may prove to be extremely useful for the study of breathy and tense voice occurring in speech data containing wide variation in expressiveness and voice quality. Specifically, this could be used to improve methods for clustering speaking styles in corpora for the purpose of building expressive speech synthesis systems (Székely et al., 2011, 2012b; Braunschweiler and Buchholz, 2011). The parameter has potential for use in discriminating speaking styles (Scherer et al., 2012b) and for studying and modelling the use of voice quality in interactive speech.

Relevant publications

- Kane, J., Gobl, C. (2013) Wavelet maxima dispersion for breathy to tense voice discrimination, *IEEE Transactions on Audio, Speech and Language processing* 21(6), pp. 1170-1179.
- Kane, J., Gobl, C. (2011) Identifying regions of non-modal phonation using features of the wavelet transform, In *Proceedings of Interspeech 2011, Florence, Italy*, 177-180.
- Scherer, S., Kane, J., Gobl, C., Schwenker, F., (2013) Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification, *Computer Speech and Language* 27(1), pp. 263-287.
- Székely, É, Kane, J., Scherer, S., Gobl, C., Carson-Berndsen, J., (2012) Detecting a targetted voice style in an audiobook using voice quality features, *Proceedings of ICASSP, Kyoto, Japan,* 4593-4596.
- Scherer, S., Layher, G., Kane, J., Neumann, H., Campbell, N., (2012) An audiovisual political speech analysis incorporating eye-tracking and perception data, *Proceedings of LREC, Istanbul, Turkey,* 1114-1120.

Chapter 8

Detection of creaky voice

Summary

This chapter describes a new algorithm for automatically detecting creaky voice in speech signals. Detection is made by utilising two acoustic parameters which are designed to characterise creaky excitations following previous evidence in the literature combined with new insights from observations in the current work. In particular the new method focuses on features in the Linear Prediction (LP) residual signal including the presence of secondary peaks as well as prominent impulse-like excitation peaks. These parameters are used as input features to a decision tree classifier for identifying creaky regions. The algorithm was evaluated on a range of read and conversational speech databases and was shown to clearly outperform the state-of-the-art. Further experiments involving degradations of the speech signal demonstrated robustness to both white and babble noise, providing better results than the state-of-the-art down to at least 20 dB signal to noise ratio.

8.1 Introduction

Several recent studies in the literature have been devoted to the development of improved methods for modelling aspects of the glottal source and excitation characteristics in speech (e.g., Drugman et al., 2009a; Degottex et al., 2011b; Kane and Gobl, 2011). Many of these aspects contribute significantly to the perception of voice quality. The focus of this chapter is on one particular voice quality sometimes referred to as creaky voice. Several voice quality labels such as glottal fry, vocal fry, laryngealisation or creak are often used in the literature and are in the vicinity of each other in terms of their physiological and acoustic characteristics. For the present chapter these labels are subsumed into one voice quality class, *creaky voice*, which will be defined solely on the basis of the auditory criterion: 'a rough quality with the sensation of additional impulses' (as is done in Ishi et al., 2008b). This approach is justified as previous studies have demonstrated some of these voice quality variations to be perceptually similar (Gerratt and Kreiman, 2001). Laver (1980), however, makes the distinction between creak and creaky voice, stating that creaky voice is a compound of creak and modal voice. This distinction is not made in the present chapter and there is evidence to suggest that such a distinction is not utilised by speakers for any linguistic or paralinguistic contrast (Ishi et al., 2008b; Laver, 1994).

Details of the physiology and the acoustics related to creaky voice are relevant to the design of the parameters described in this chapter. The reader can refer to Chapter 2, Section 2.1.2 for these details.

tract resonances have almost completely decayed before the start of the next pulse.

The distinctive acoustic characteristics of creaky voice can cause problems for standard speech analysis methods (including f_0 tracking and spectral analysis). The very low f_0 values and, often, irregular temporal patterning may not be properly handled by standard f_0 tracking algorithms. Standard frame lengths (usually no longer than 32 ms) may be too short to capture two glottal pulses and, hence, will be unsuitable for obtaining strong periodicity information. Commonly used f_0 trackers tend to either output spurious values in creaky regions or consider creaky regions to be unvoiced. As a result of this, creaky regions will be poorly modelled in most speech technology applications. This problem was highlighted in a previous study (Silen et al., 2009) which involved the development of a speech synthesis system for Finnish (a language in which creaky voice frequently occurs).

However, creaky voice is commonly produced for a range of interactive, expressive and stylistic reasons. It has previously been studied in relation hesitations (Carlson et al., 2006) and turn-taking (Ogden, 2001), as well in the context of various forms of expression and emotion (Yanushevskaya et al., 2005; Gobl and Ní Chasaide, 2003b; Ishi et al.,

2008a). It has also been shown to frequently occur at phrase boundaries and utterance final position in American English (Surana and Slifka, 2006b). If low subglottal pressure is a necessary condition for creak, then it is not surprising that it frequently occurs in phrase/utterance/turn-final position, where the speaker will have less air available than at the start of the utterance. Creaky voice has also recently received attention from popular science articles following the study in Wolk and Abdelli-Beruh (2012) which demonstrated that two thirds of the young female American English speakers analysed, displayed creaky voice at the end of read sentences. The authors state that continuous use of creaky voice is likely to be more prevalent in more sociable, conversational speech settings (however this was not formally investigated).

The study of creaky voice (and indeed voice quality in general) has been hindered because of the lack of suitable automatic detection algorithms and, as a result, most applied studies on creaky voice tend to either rely on qualitative interpretation or use small amounts of data.

In terms of specific applications robust detection could be used to segment creaky regions in corpora used for text-to-speech synthesis which would facilitate the use of more appropriate acoustic modelling and, hence, better rendering of these regions (Drugman et al., 2012b). This would be particularly important for expressive or conversational speech synthesis as the use of creaky voice (and indeed other aspects of voice quality) are known to play a critical role in spoken interaction (Campbell and Mokhtari, 2003), e.g., with turn-taking (Ogden, 2001). Also, as creaky voice is known to be frequently produced during hesitations (Carlson et al., 2006) its detection could also be used to help identify hesitations which could, in turn, be used for distinguishing speaking styles or, for instance, providing feedback on presentation skills. The robust automatic detection of creaky voice would be beneficial for sociological studies (e.g., Wolk and Abdelli-Beruh, 2012) and studies on tonal patterns (Yu and Lam, 2011) in terms of allowing quantitative analysis on larger volumes of data. Furthermore, as studies have shown listeners to be sensitive to creaky voice in terms of recognition of the speaker's identity (Böhm and Shattuck-Hufnagel, 2007), the detection of creaky voice can be exploited for improving speaker recognition systems (Espy-Wilson et al., 2006; Elliot, 2002).

Motivated by this, the present chapter describes a new algorithm for automatically detecting creaky voice through the use of two acoustic parameters which describe aspects of the LP-residual signal. The approach builds on previous work the present author has been involved in (Drugman et al., 2012a). This initial study involved the use of a single parameter which was evaluated on a rather small set of read text-to-speech (TTS) synthesis data. The current chapter involves the inclusion of a further acoustic parameter

as well as the incorporation of the features into a classifier for detecting creak. A much larger evaluation is carried out here on a wide range of speech data, covering a variety of speakers, gender, languages, recording conditions and speaking styles. Furthermore, additional robustness experiments are conducted examining the effect of different noise types and levels on the performance of the different methods.

8.2 State-of-the-art

Although a considerable amount of research has been carried out investigating the acoustic characteristics of creaky voice, there is a clear lack of algorithms for detecting it automatically. Some studies describe automatic detection of 'irregular phonation' (see e.g., Böhm et al., 2010; Surana and Slifka, 2006a; Vishnubhotla and Espy-Wilson, 2006), a class within which creaky voice is contained. For instance in Böhm et al. (2010) the authors derive six acoustic parameters and use them as input to a support vector machine (SVM) based classification system. Their method involved using acoustic measurements from previous studies (i.e. Surana and Slifka, 2006a; Ishi et al., 2008b). In Surana and Slifka (2006a) the authors propose the use of acoustic parameters including normalised root mean squared amplitude and smoothed-energy-difference amplitude measures. However, misdetections apparently occur in low f_0 regions.

In the present chapter two creaky voice detection algorithms from the literature are included (Ishi et al., 2008b; Vishnubhotla and Espy-Wilson, 2006) for comparison with the proposed algorithm. They are now described in detail.

8.2.1 Ishi's method for detection of vocal fry/creak

Recently an algorithm was presented for the automatic detection of vocal fry/creaky voice (Ishi et al., 2008b) which builds on previous work by the same authors (Ishi, 2004; Ishi et al., 2005). This algorithm involves detecting candidate regions in a power contour measured from a bandlimited speech signal. Then a combination of autocorrelation and cross-correlation methods are used to discriminate creaky voice from 'normal' voiced speech and unvoiced/silence regions, respectively. The full details of the algorithm are as follows.

The algorithm operates on the speech signal, which has been bandlimited to 100 - 1500 Hz. A 'very short-term' power contour is measured, with a frame length of 4 ms and shift of 2 ms, in order to highlight the amplitude variation within individual pulses (see Figure 8.1 panel b). Peaks are then detected in this contour and Power Peak (PwP) parameters are derived for each peak based on the previous (PwP-rising) and following

(PwP-falling) 5 frames (i.e. 10 ms) in the contour. The maximum power difference in each direction is used as the PwP value and a threshold is applied to this parameter to determine whether the peak can be used as a creak candidate location. In addition to this, it has been suggested by the author (personal communication) that peaks more than 20 dB below the maximum power peak (for each utterance) can also be discarded.

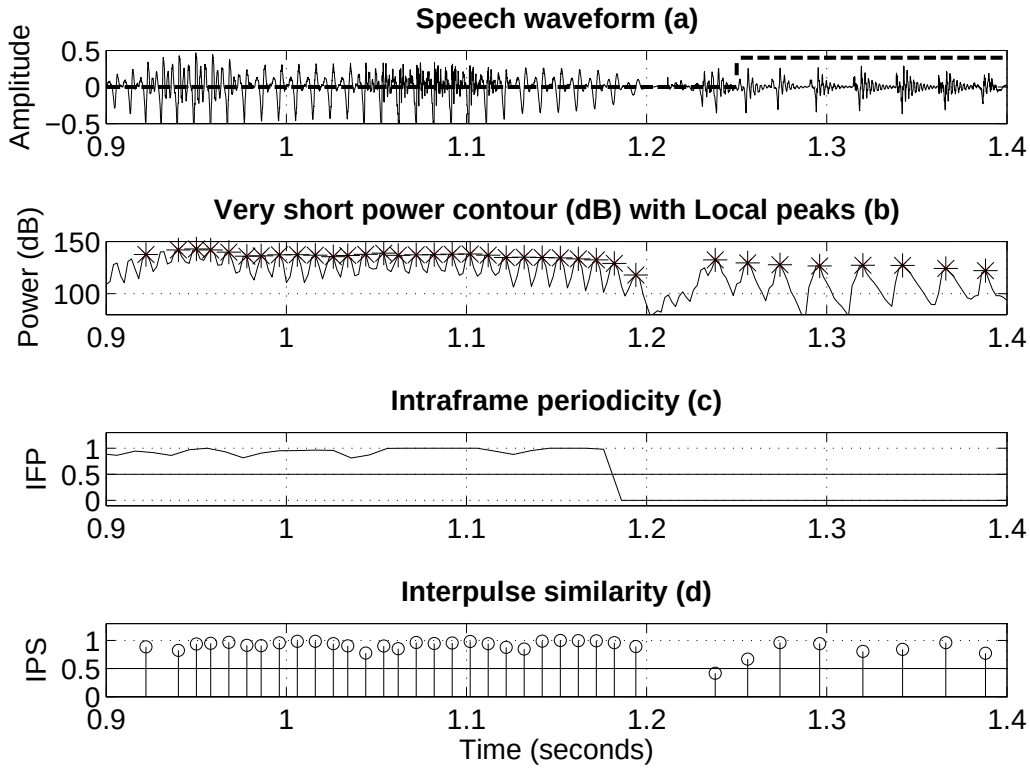


Figure 8.1: Illustration of creaky voice detection using the method proposed in Ishi et al. (2008b). The speech waveform, with binary creak decision (dashed line), is shown with the creaky region beginning from around 1.22 seconds (panel a), along with the very short term power contour and detected local peaks (panel b), the Intraframe Periodicity (IFP) contour (panel c) and the Interpulse Similarity (IPS) values (panel d). Horizontal lines for IFP and IPS are given to illustrate the thresholds used.

Given these peak locations (creak candidates) a check is performed against a frame-synchronised periodicity strength measure in order to discriminate creaky regions from 'normal' voiced regions. The Intra-Frame Periodicity (IFP, see Figure 8.1, panel (c)) contour is calculated with:

$$\text{IFP} = \min\left\{\frac{N}{N-\tau}\cdot\text{autoCorr}(\tau);\quad \tau = j\cdot\tau_0;\quad j = 1,2,...\right\} \tag{8.1}$$

where N is the frame length (set to 32 ms), τ is the autocorrelation lag, autoCorr is

the normalised autocorrelation function, and τ_0 is the lag of the strongest autocorrelation peak. Note also that the search space for τ is limited to 15 ms and that the factor $\frac{N}{N-\tau}$ is used to compensate for the decrease in amplitude with increasing τ in the autocorrelation function. 'Normal' voiced regions are expected to have an IFP value close to 1, while creaky (and non-voiced) regions are likely to show a value closer to 0. This is based on the observation that creaky regions can display irregular temporal patterns and also, that even in cases where creaky regions display a reasonable amount of periodicity, the very long pulses will mean that the frame is not sufficiently long to capture strong periodicity information. Finally, IFP values are set to 0 unless three successive frames are found to be above a given threshold.

Next an Inter-Pulse Similarity (IPS, see Figure 8.1, panel (d)) measure is calculated with:

$$\text{IPS} = \max\left\{\text{CCorr}(F_{\tau_1}, F_{\tau_2}); \quad \tau_1 - \tau_2 < T_{max}\right\} \tag{8.2}$$

where CCorr is the cross-correlation function, F_{τ_1} and F_{τ_2} are the frames centred on successive candidate peak locations, and T_{max} is the maximum allowed distance between adjacent peaks, and is set to 100 ms. Each frame is selected as the range of 5 ms around the peak location. Adjacent creak pulses are expected to display a reasonably high similarity (as the vocal tract setting is not expected to have significantly changed) and, hence, IPS values are expected to be high (e.g., above 0.5). Non-speech and unvoiced regions, on the other hand, are expected to display low levels of similarity and, hence, and IPS close to 0. If adjacent pulses are too far apart, i.e. above T_{max}, IPS values are also set to 0.

The optimal thresholds suggested in the original publication (Ishi et al., 2008b) are used, i.e. $\{\text{PwP} \geq 7 \text{ dB} \;\&\; \text{IFP} \leq 0.5 \;\&\; \text{IPS} \geq 0.5\}$ for a peak to be considered to be creaky (however, different thresholds have also been applied in a separate applied study, Ishi et al., 2008a). The binary creak decision is then made by merging regions between detected creak peaks (see Figure 8.1, panel (a)). This method is given the label **Ishi Orig.** throughout this chapter.

In the original publication (Ishi et al., 2008b) the authors report an upper bound detection rate of 74 % with a false alarm rate of 10 %, using thresholds optimised on the same dataset. Note that frame level results were not reported.

8.2.2 Extension of the Aperiodicity, Periodicity and Pitch (APP) detector

This method has been proposed in Vishnubhotla and Espy-Wilson (2006) for the automatic detection of 'irregular phonation' (a term they say they use interchangeably with creak). The authors interpret this label also to include sounds referred to as vocal fry, diplophonia, glottalisation, laryngealisation, pulse register phonation and glottal squeak (Vishnubhotla and Espy-Wilson, 2006).

As part of the algorithm they make use of the Aperiodicity, Periodicity and Pitch (APP) detector (originally presented in Deshmukh et al., 2005). This involves applying a gamma-tone filterbank to decompose the speech signal into 60 frequency bands. An Average Magnitude Difference Function (AMDF) is calculated on the separated frequency bands (smoothed by the use of the Hilbert envelope) to determine aperiodicity/periodicity in the signal. The AMDF function, $\gamma_n(k)$, for each outputted signal is calculated with:

$$\gamma_n(k) = \sum_{m=-\infty}^{\infty} |x(n+m)w(m) - x(n+m-k)w(m-k)| \tag{8.3}$$

where $w(m)$ is a rectangular window centered on n and given a specified width. When the signal is periodic the AMDF function will display pronounced 'dips' when k (the lags of the function) is equal to integer multiples of the fundamental period. In the implementation of the algorithm which was obtained, the frame length was set to 25 ms, with a shift of 2.5 ms.

Next, 'irregular phonation' is differentiated from aperiodic frames, breathy vowels and voiced fricatives using the so-called dip profile of their AMDF in various frequency bands. The dip profiles for 'irregular phonation' display distinctive clustering characteristics to both regular phonation and speech with turbulent excitations. Identification of 'irregular phonation' is on the basis of detection of this characteristic.

Finally, the problem of false positives in some stops is addressed by calculating the spectral slope. This is done by fitting a regression line to the amplitude spectrum from 2 to 4 kHz. A threshold of -0.05 is empirically used to differentiate 'irregular phonation' from these stops. In the evaluation, the original implementation of the algorithm was used, kindly shared by the authors. The method is given the label **Vishnu**.

8.3 Proposed method

The current section presents a description of the proposed algorithm for automatically detecting creaky regions in a speech signal. The method is based on further development of the algorithm described in Drugman et al. (2012a). First an analysis of the excitation characteristics of creaky voice is carried out which highlights the main acoustic features which the algorithm is designed to detect. Then follows a description of the two components of the algorithm, both of which are used as input features to a binary classifier (see Section 8.3.4). Note that the two components attempt to describe different aspects of the LP-residual signal in creaky regions. These two aspects may, at times, both be present. At other times, however, just one is displayed.

8.3.1 Excitation characteristics

Following extensive qualitative analysis of the Linear Prediction (LP) residual signal (obtained by LPC analysis and subsequent inverse filtering) some distinctive characteristics were observed which appeared to be closely associated with creaky regions. The first observation was the presence of secondary (or even tertiary) peaks proceeding the main excitation peak which corresponds to the glottal closure instant (GCI, Drugman et al., 2012c). This is illustrated in Figure 8.2, where strong residual peaks can be observed before the residual peak which corresponds to the GCI (as shown by the derivative of the EGG signal). Although secondary excitations and double-pulsing have frequently been reported in the literature (e.g., Blomgren et al., 1998; Gobl and Ní Chasaide, 1992) these secondary residual peaks frequently did not appear to correspond to secondary laryngeal excitations (which would show up in the EGG signal). Instead, it is hypothesised that often these peaks correspond to an abrupt glottal opening following a long closed phase. Some preliminary comparison with glottal source derivative signals, estimated by inverse filtering, and EGG signals appeared to support this. Nevertheless, strong secondary laryngeal excitations did at times cause secondary peaks to occur in the LP-residual signal.

Following these observations, Component 1 of the algorithm is designed to exploit these secondary peaks occurring in the LP-residual signal.

Further analysis of the LP-residual of creaky speech regions revealed that although the above trend is very prevalent, secondary LP-residual peaks may sometimes be absent. Considering the LP-residual signal in Figure 8.3 one can observe strong impulse-like peaks with no secondary peaks, even when the EGG derivative is displaying small secondary peaks.

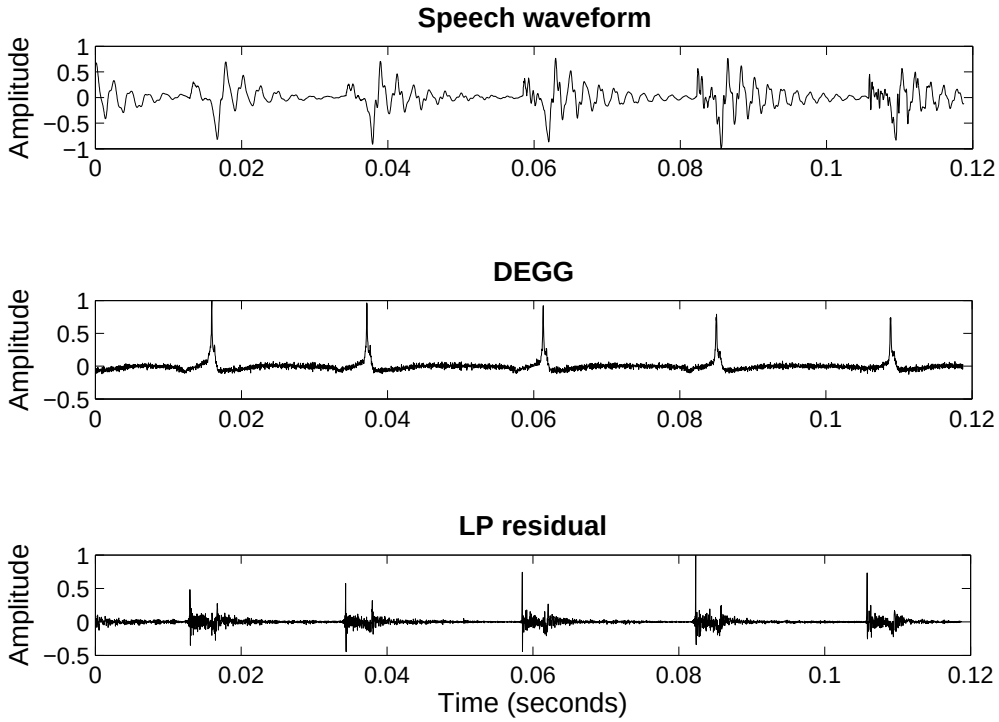


Figure 8.2: Speech waveform (top panel), EGG derivative (middle panel) and LP-residual (bottom panel) signals from a creaky region in an utterance produced by a male speaker.

Component 2 of the algorithm is, hence, designed to capture this specific feature combined with the knowledge that creaky voicing produces considerably longer glottal pulse duration (Titze and Sundberg, 1992; Blomgren et al., 1998).

8.3.2 Component 1: Detection of secondary excitation peaks

The block diagram of Component 1 of the proposed algorithm is shown in Figure 8.4. This method is designed to exploit the secondary peaks in the residual excitation signal. The residual signal is estimated following LPC analysis. The aim of this analysis is to cancel the spectral contribution of both the vocal tract filter and the glottal source signal, thereby rendering the spectrum of the LP-residual essentially flat. However, this residual signal exhibits important phase properties of the excitation source, including the presence of secondary peaks. The key idea of Component 1 is that when applying a resonator to the LP-residual, secondary peaks will perturb its output and produce a greater amount of harmonics. The full details of this method are as follows.

The LP-residual is obtained by LPC analysis of order $(fs/1000) + 2$ [1], where fs is

[1]This LPC order is chosen in order to obtain a spectral envelope which is not biased to harmonics

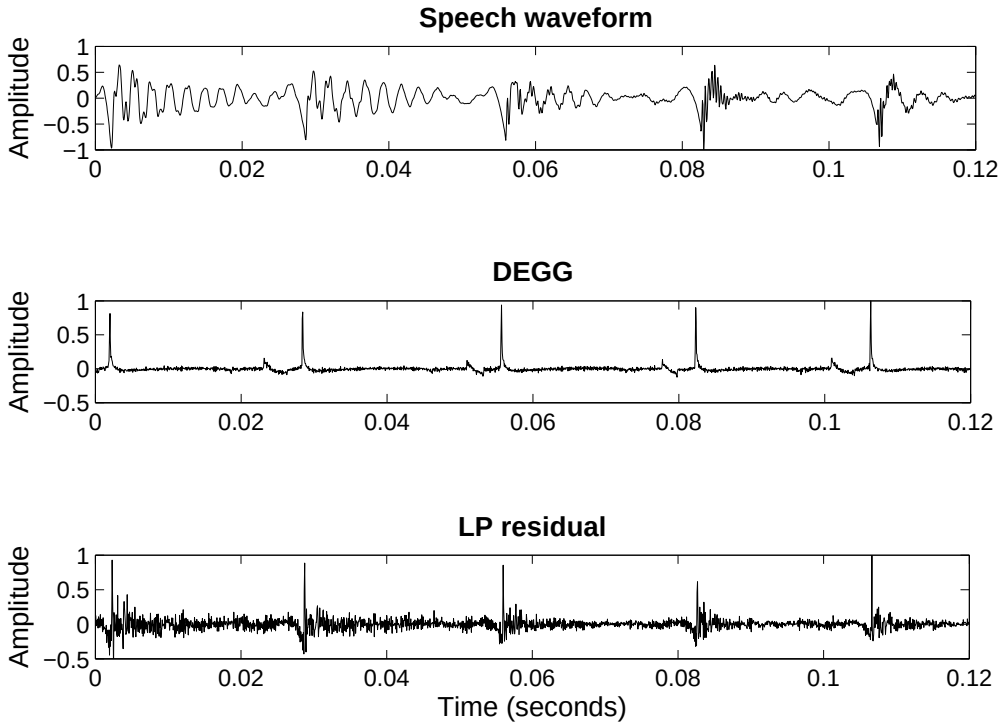


Figure 8.3: Speech waveform (top panel), EGG derivative (middle panel) and LP-residual (bottom panel) signals from a creaky region in an utterance produced by a male speaker.

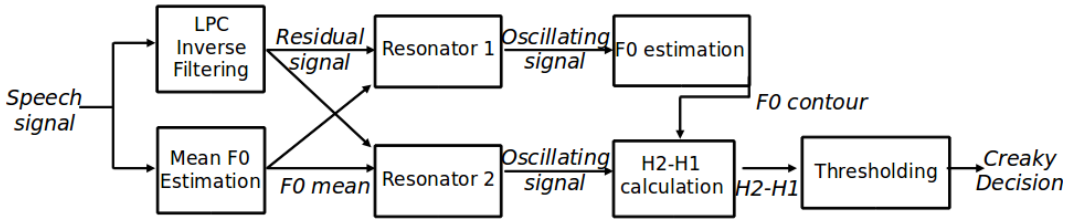


Figure 8.4: Workflow of Component 1 of the proposed technique. The H2-H1 parameter is derived from the response of two resonators excited by the LP-residual signal. Both resonators are centred on $f_{0,mean}$ but use different bandwidths. For details see text.

the sampling frequency. Two separate resonators are then applied to the residual signal for different purposes. One is used for getting a more robust estimate of the f_0 contour in creaky regions while the other is to highlight the presence of secondary residual peaks. Both resonators use a centre frequency of $f_{0,mean}$, the mean f_0 of the speaker. This value is here estimated using the Summation of Residual Harmonics (SRH) method (Drugman et al., 2011), although the choice of algorithm for this is not critical. Both resonators are

and is frequently used in the literature. Although it is widely known that LPC analysis becomes biased towards harmonics for high f_0 values (Villavicencio et al., 2006; Kay, 1988), creaky regions display a low f_0 and here LPC analysis was found to be suitable.

characterised by two complex conjugate poles.

For estimating the f_0 contour the bandwidth of Resonator 1 was set to 1000 Hz as it gives a reasonable compromise between avoiding ambiguity with octave jumps (bandwidth too high) and capturing the spread of f_0 values from the $f_{0,mean}$ often found in creaky parts (which might not be achieved correctly if the bandwidth is too low). To estimate the local f_0, a 50 ms-long Hanning window is applied to the resonator output and the corrected autocorrelation function $r'(\tau)$ is calculated:

$$r'(\tau) = \frac{N}{N-\tau} \cdot \text{autoCorr}(\tau) \tag{8.4}$$

where N is the window length (in samples) and τ is the number of autocorrelation lags. A correction of $\frac{N}{N-\tau}$ is applied to compensate for the decreasing properties of the autocorrelation function with increasing τ (as is used in Ishi et al., 2008b). The local fundamental period is then considered as the position of the maximum value in $r'(\tau)$ above the peak centred on $\tau = 0$.

For highlighting secondary excitations, a more pronounced resonating character is needed and, hence, the bandwidth of Resonator 2 is set to 150 Hz. To measure the importance of secondary pulses, the amplitude difference (in dB) between the two first harmonics ($H2 - H1$) is computed on the spectrum of the autocorrelation function, as it allows to enhance harmonic peaks. Note that $H2 - H1$ is then filtered by a 100 ms-long moving average filter to lessen the impact of outlier values.

An example of the contour is shown for a speech utterance in Figure 8.5 where it is clear that H2-H1 reaches much higher values in the annotated creaky voice region at the end of the speech segment, and for which applying a threshold (of around 0 dB in this case) would lead to good detection results.

An illustration of the steps involved in the workflow (depicted in Figure 8.4) is given in Figure 8.6 for both a segment of speech involving modal phonation (on the left) and creaky voice (on the right). The speech signal is displayed in the top row plots. In the middle row, the residual signal and the output of Resonator 2 are represented.

In the case of modal phonation, it can be noted that the residual signal exhibits a regular structure with major peaks only at the GCI positions. As a result, perturbations between two major excitation peaks are relatively weak and the oscillating signal outputted by Resonator 2 will only contain a small amount of harmonics. This is reflected in its amplitude spectrum (in dB) in the bottom row of Figure 8.6 where the level at f_0 is much higher than for the second harmonic (i.e. $2 \cdot f_0$). The difference H2-H1 in such a modal phonation then reaches low negative values (-15 dB in the case of Figure 8.6).

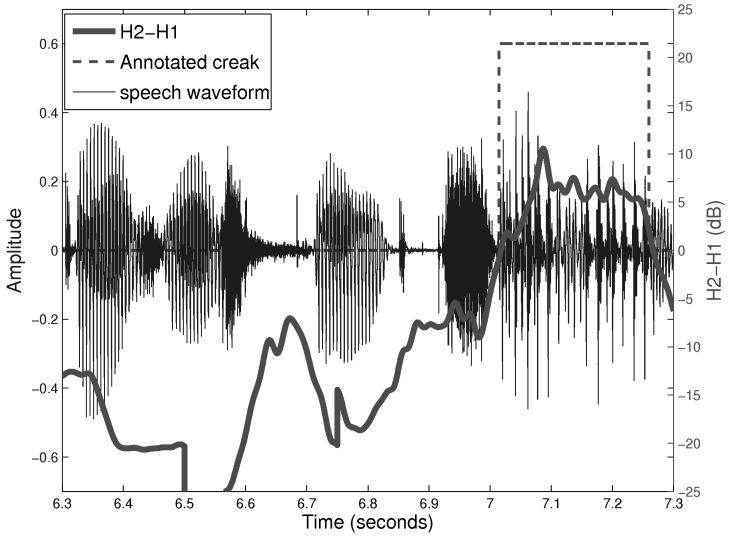


Figure 8.5: Illustration of the H2-H1 contour (thick line) for a speech segment with annotated creaky voice region (dashed line). The left y-axis shows the scale for the amplitude of the speech signal, whereas the right y-axis shows the scale for the H2-H1 contour in dB.

On the other hand, for creak, secondary pulses significantly re-excite the resonator between two consecutive GCIs, leading to perturbations in its output. This effect is seen in the corresponding amplitude spectrum which displays a greater richness of harmonics. More specifically, the level at the second harmonic is much higher compared to the modal phonation case, and can even exceed the level at f_0. As a consequence, the presence of secondary peaks in the excitation of creaky voice is reflected by a higher harmonicity in the output of Resonator 2. This leads therefore to higher values of H2-H1 (9.22 dB for the creaky example of Figure 8.6) compared to what is obtained for modal phonation.

8.3.3 Component 2: Residual peak prominence

Component 2 of the algorithm is designed to detect creaky speech regions where there are prominent residual peaks (as shown in Figure 8.3). The prominent residual peaks may stem from the sharp vocal fold closure resulting from high levels of adductive tension (Laver, 1980). This is combined with the knowledge that creaky regions contain very long glottal pulses (Blomgren et al., 1998; Gobl and Ní Chasaide, 1992). The very long glottal pulses may be due to ventricular incursion (Edmondson and Esling, 2006) where the ventricular folds push down on the *true* vocal folds, causing an increased mass which vibrates at a lower frequency. Creaky regions can, at times, display irregular temporal patterning which can render frequency domain methods unsuitable. It follows that this second component of the algorithm does not rely on signal information to do with periodicity but instead

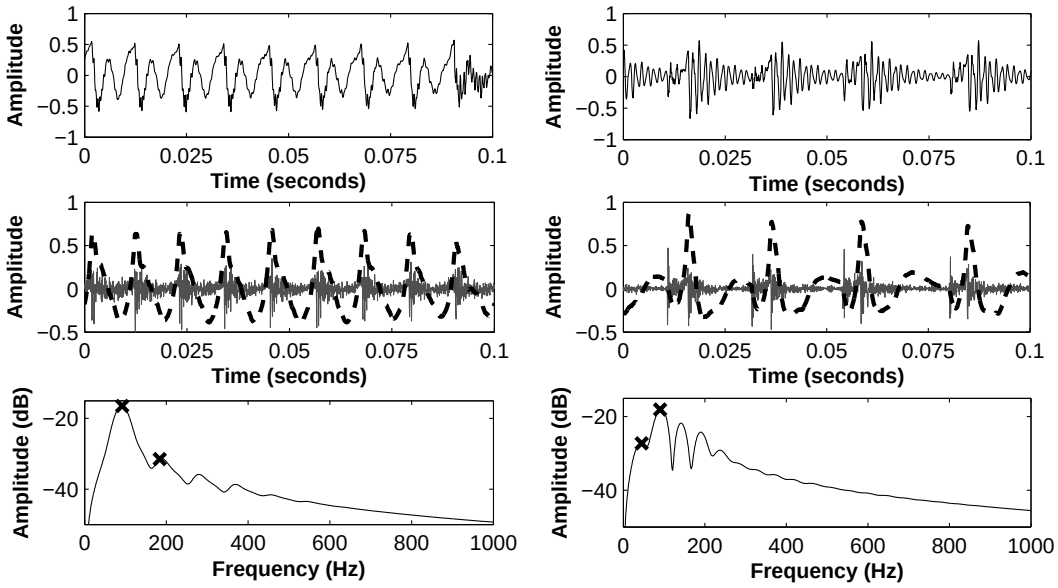


Figure 8.6: Example of modal phonation (left column) and creaky voice (right column) uttered by the same speaker, with the speech waveform (top row), the LPC residual signal (middle row, solid line) together with the output of Resonator 2 (in dashed line), and the amplitude spectrum (in dB) of a frame of the output of Resonator 2 where the values for f_0 and $2 \cdot f_0$ are indicated by crosses (bottom row).

looks to characterise individual pulses in the time domain. The method is carried out as follows.

Initially residual peak prominence was measured directly from the LP-residual signal, but further examination revealed this approach to be rather sensitive to additive noise. Instead, the output of Resonator 1 (see Figure 8.4) was used. The setting of the resonator bandwidth to 1000 Hz is suitable for highlighting the prominence of the main residual peaks without being overly biased towards secondary peaks (if present).

The method operates on a fixed, non-overlapping frame basis using a rectangular window and with a frame-length of 30 ms. This roughly corresponds to two periods at 70 Hz. In this method correct polarity of the speech signal is assumed (this can be determined automatically for example using the method described in Drugman and Dutoit, 2011) and the output of the resonator will display strong negative peaks. This signal is inverted so that it displays strong positive peaks, corresponding to positive peaks in the LP-residual.

For each frame the absolute maximum peak in the resonator output is identified and the frame is then shifted to be centred on this peak. Considering Figure 8.7 one can observe for a creaky region (right panel) the prominent peak amplitude in the centre of the frame. For a modal region (left panel), however, peaks from neighbouring glottal

pulses are captured within the frame length.

By measuring the amplitude difference between the maximum peak (in the centre of the frame) and the next strongest peak one can obtain a parameter value which differentiates modal and creaky regions. In order to avoid selecting values in the vicinity of the main centre peak, the search for the next strongest peak is made outside a distance of 6 ms of the centre of the frame. This corresponds to 40 % of half the frame length which ensures that there is sufficient space for peaks to occur from neighbouring glottal pulses. A value is thus obtained for each frame producing the outputted parameter contour. This contour was then filtered with a 3-point median filter to remove misdetections due to transients in the signal.

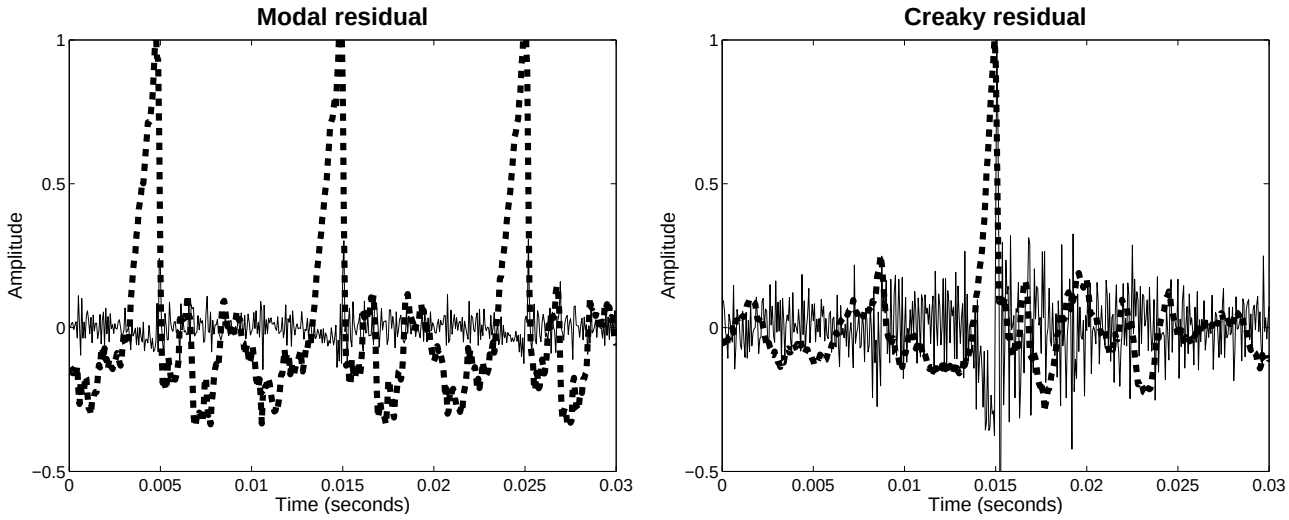


Figure 8.7: A 30 ms peak centred residual frame (thin line), with superimposed Resonator 1 output (thick dashed line) for a modal (left panel) and a creaky (right panel) utterance segment. The residual peak prominence value for the modal segment is very close to 0, while the value for the creaky segment is over 0.75. Both LP-residual and resonator output signals are normalised in amplitude for clarity.

8.3.4 Classification using the two parameters

In order to detect creaky regions the two parameters were used as input features to a binary decision tree classifier (Breiman et al., 1984). The separation of the two classes (i.e. creak and non-creak) is done using a top-down approach where both classes are initially placed at the root node and then a series of binary questions are asked (to do with the input features) and for each question a new child node is created. This creates the decision tree, the ends of which are leaf nodes.

Decision trees were developed on a training set using the extracted parameter values and the binary creak annotation labels (see Section 8.5.1). When inputting a training example to the trained classifier the output is the posterior probability, P_1, of the example corresponding to class 1 (i.e. creaky) and the posterior probability, P_0, of it corresponding to the class 0 (i.e. non-creak). The standard binary decision is typically set to 1 if $P_1 > 0.5$, and otherwise 0. In the training phase the decision tree classifier is optimised for minimising the error rate. However, for skewed datasets which contain a given class to be detected which displays sparse occurrence (e.g., creaky voice or laughter) this error criterion is not suitable. To address this a further processing was carried out during the training phase. This involved making the decision strategy such that if $P_1 > \alpha$ then the output is set to 1, otherwise it is set to zero. α is varied in the range [0, 1] and the setting which produced the higher F1 score on the training set is subsequently used for the decision strategy in testing. Note that the F1 score is a more suitable criterion for a dataset of the type used in this study (see Section 8.5.2).

In this chapter four methods involved the use of the classification approach described above. The first component (labelled **Comp. 1**) and the second component (labelled **Comp. 2**) of the proposed method were used separately and in combination (labelled **Comp. 1 & 2**). The four parameters derived using the method by Ishi et al. (2008b) and described in Section 8.2.1 were also included, i.e. PwP-rising, PwP-falling, IFP and IPS (this method was labelled **Ishi Opt**).

8.3.5 Post-processing

Some final post-processing can then be carried out on the binary decision vector of the proposed (or other) methods. To help remove misdetections in unvoiced and non-speech areas, zero-crossings are measured on 20 ms frames. Areas with a zero crossing rate (ZCR, i.e. number of zero-crossings per ms) of more than 5, were considered to be unvoiced or silent parts and therefore excluded as potential creaky voice areas. Note that the use of energy contours was deemed unsuitable particularly because conversational speech data can display widely varying energy values. This is, of course, a rather basic method for determining unvoiced regions and could be substituted with a more sophisticated method (see for example Ghosh et al., 2011).

Finally, overly short detected creak regions were removed and nearby adjacent creak regions were merged. A minimum creak length of 30 ms was used which corresponds roughly to two periods at 70 Hz. The assumption here is that at least two pulses are required for the perception of creaky voice and, again, that the perception of individual

glottal pulses starts around 70 Hz (Titze, 1994). The binary creak decision vectors used in this study were sampled every 10 ms, so the removal of short regions and merging of close regions was done in the same operation through the use of a 5-point (i.e. 50 ms) median filter applied to the binary decision vector. For instance this would remove one or two positive creaky voice samples (i.e. ones) surrounded by negative decision samples (i.e. zeros). On the other hand, if a zero had two positive decision samples on either side, the median filter would merge the two regions and convert the zero to a one.

8.4 Speech material

In order to provide a thorough evaluation of detection performance a wide range of speech databases was used which covered gender, language, read/conversational speech and recording condition variations. All speech data were downsampled to 16 kHz. A summary of the speech data used in the evaluation in the present chapter is given in Table 8.1.

8.4.1 Text-To-Speech databases

To evaluate performance in *ideal* recording conditions 100 sentences containing creaky regions were selected from three Text-To-Speech (TTS) databases. An American male speaker (BDL) was selected from the ARCTIC database (Kominek and Black, 2004). Utterances from a Finnish male speaker (MV, as was used in Vainio, 2001) and from a Finnish female speaker (HS, as was used in Silen et al., 2009) were also used.

8.4.2 Spontal corpus

Next the author wished to include conversational speech data recorded in high quality conditions. The Spontal corpus (described in Edlund et al., 2010) contains audio, video and motion capture data from a large number of dialogues lasting at least 30 minutes, carried out in a recording studio. The dialogues were in Swedish and participants were encouraged to talk about whatever topic they wished.

The audio data from the microphone channel of one male (label: 09-13-02) and one female speaker (label: 09-03-01) was selected, as these speakers were deemed to produce frequent creaky utterances. Audio was captured through the use of two microphones per speaker: a Brüel & Kjær 4003 omni-directional goose-neck at 1m distance, and a head-mounted Beyerdynamic Opus 54 cardioid which was used to obtain optimal recording

quality. In the current chapter only the audio channels from the head-mounted microphone were used. The original sampling rate was 48 kHz.

8.4.3 American conversational data

Audio streams were selected from two male and two female American English speakers who engaged in conversations on the topic of food. The speech data were part of a larger number of conversational recordings, additional to the data used in Yuasa (2010). For each speaker the audio was of approximately 10 minutes in duration and was captured using a headset resting around the speaker's neck with the microphone pointing to the mouth. Conversations were carried out in a booth of a media center, a relatively quiet but relaxing environment suited for natural conversations.

8.4.4 Japanese conversational data

Also included were the audio recordings of two female Japanese speakers, previously described in Magnuson (2011). Both speakers spoke a Japanese dialect spoken in Western Japan and engaged in a 30 minute conversation. The speakers were shown some short animated films before starting the conversation. For the conversation itself they were encouraged to talk about any topic they wished but to refer to the short film at some stage during the conversation. Audio was recorded on AKG C420 III PP MicroMic headset microphones wired through a BeachTek DXA-2S pre-amp connected to the video camera (Sony DCR-TRV38 Mini DV camera). WAV files were extracted from the video into separate channels.

8.5 Experimental setup

For the experiments conducted as part of this study creaky voice regions were detected using the proposed algorithm (Section 8.3) as well as Ishi's algorithm (Section 8.2.1) and the algorithm by Vishnubhotla (Section 8.2.2). The experimental setup is now described in full.

8.5.1 Human annotation

Unfortunately there is no obvious way of obtaining an automatic reference for creaky regions in speech. Furthermore, the only relatively large database of speech data labelled for creaky voice was the database used in Ishi et al. (2008b). As a result, in order to have

Table 8.1: Summary of the speech data used for evaluating creaky voice detection performance.

Database	ID	Gender	Country	Conditions	Speech	Duration
TTS	US-M	Male	USA	Studio	Read	100 Sentences
	Fin-F	Female	Finland	Studio	Read	100 Sentences
	Fin-M	Male	Finland	Studio	Read	100 Sentences
Spontal	F	Female	Sweden	Studio	Conversation	30+ Minutes
	M	Male	Sweden	Studio	Conversation	30+ Minutes
US	F1	Female	USA	Quiet room	Conversation	10+ Minutes
	M1	Female	USA	Quiet room	Conversation	10+ Minutes
	M2	Male	USA	Quiet room	Conversation	10+ Minutes
	F2	Male	USA	Quiet room	Conversation	10+ Minutes
Japan	F1	Female	Japan	Quiet room	Conversation	30+ Minutes
	F2	Female	Japan	Quiet room	Conversation	30+ Minutes

a reference to evaluate detection performance human annotations of the speech data was carried out. At the same time the aim was to evaluate performance of the algorithms on a large set of data covering a range of different speaking styles, languages, recording conditions, etc. As the manual annotation of this volume of data is both tedious and time-consuming, a single person (the present author) carried out the annotation. The annotator strictly followed the annotation procedure outlined in Ishi et al. (2008b). Ultimately the binary decision on the presence of creaky voice was based on the auditory criterion "a rough quality with the additional sensation of repeating impulses". However, the annotation was also guided through the use of spectrograms and f_0 contours. Wideband spectrograms typically display vertical striations (Ogden, 2009) and f_0 contours can frequently display spurious values or disappear (i.e. are considered unvoiced) and, hence, these displays were used to help guide the annotation.

Furthermore, manual voice activity segmentation was carried out for the speech data containing conversational speech. The exception to this was in the Spontal corpus where automatic voice activity detection was carried out with the algorithm proposed in Heldner et al. (2011).

The percentage of time speaking which was annotated as creaky voice was calculated for the 11 speakers used in the evaluation and the average was 6.7 % with a range of 3.6 - 10.5 %.

8.5.2 Evaluation metrics

To assess the performance of the algorithms evaluation metrics were calculated at both the event level and the frame level. For the event level the metrics used were: hit (i.e. some part of a reference creak region was correctly detected), miss (i.e. for a reference creak region no positive detection was made) and false alarm (i.e. within a detected creak region there was no reference creak).

At the frame level the standard metrics were used, i.e. True Positive Rate (TPR, also known as recall):

$$TPR = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{8.5}$$

and False Positive Rate (FPR):

$$FPR = \frac{\text{False positives}}{\text{False positives} + \text{True negatives}} \tag{8.6}$$

Note that True Positive refers to a given frame containing creaky voice that has been detected as containing creaky voice. False Negative is where again the frame contains creaky voice but it has not been detected. For False Positive the frame does not contain creaky voice but has been detected as containing it. Finally, True Negative refers to a non-creaky frame which has not been detected as containing creaky voice. The F1 score is also used which combines true positives, false positives and false negatives into one single metric. This metric is particularly useful when analysing skewed datasets where the feature being considered has a rather sparse occurrence (e.g., for laughter detection, Scherer et al., 2009), and is therefore well suited for assessing the performance of creaky voice detection techniques. The metric is bound between 0 and 1, with 1 indicating perfect detection:

$$F1 = \frac{2 \cdot \text{True positives}}{2 \cdot \text{True positives} + \text{false positives} + \text{false negatives}} \in [0, 1] \tag{8.7}$$

8.5.3 Experiments on clean speech

In order to evaluate the detection performance of the various methods, analysis was carried out on the speech databases described in Section 8.4. For the methods using the decision tree classifier (i.e. Comp. 1, Comp. 2, Comp. 1 & 2 and Ishi Opt.), a leave one speaker out design was used whereby the speech data of a given speaker was held out for testing and the remainder of the speech data was used for training the classifier and optimising the decision strategy (see Section 8.3.4). The procedure was repeated for each speaker.

For the methods Vishnu. (Vishnubhotla and Espy-Wilson, 2006) and Ishi Orig. (Ishi et al., 2008b) the same settings as were described in the original publications were used for all speakers.

There were three main aims of the experiments on clean speech:

1. A preliminary assessment of the suitability of each of the classification methods for detecting creaky regions.

2. To examine the effect of the post-processing (described in Section 8.3.5) on the methods, with analysis of the entire speech dataset.

3. To provide a thorough presentation of the performance of the difference methods using both event and frame level metrics.

8.5.4 Robustness to additive noise

In addition to examining detection performance on a range of speech databases experiments were also carried out to analyse the robustness of the algorithms to additive noise. For this only the TTS databases were used which were recorded in high-quality conditions.

Degraded conditions were simulated by adding noise to the original speech waveform at various signal-to-noise ratio (SNR) settings. Both white noise and babble noise (also known as cocktail party noise) were considered. The noise signals were taken from the Noisex-92 database (Varga and Steeneken, 1993). All parameters from the different methods were extracted from speech with the various types and levels of noise added. Then a similar leave one speaker out approach, as was used described in Section 8.5.3, was applied. Here, however, optimisation of the classification was done on a 'clean' training set, whereas testing was carried out on a test set with noise added. The F1 score was calculated for each test set (i.e. each speaker) for the various noise types/levels. Note that for these robustness testing experiments the zero-crossing rate (ZCR) feature used as part of the post-processing was omitted. This is because additive noise is likely to severely affect the rate of zero-crossings and, hence, this will hinder the assessment of the detection methods themselves.

8.6 Results on 'clean' data

8.6.1 Preliminary results on TTS database

An illustration of the performance of the six detection methods on the TTS databases is shown in Figure 8.8. The F1 scores for Comp. 1 and Comp. 2 are comparatively high, and the synergic effect of their combination is clearly apparent with an important improvement over the individual components for each of the three speakers. The prevalence of the two measures can be observed here and although Comp. 1 is more prevalent, Comp. 2 still provides better detection than the other comparison methods for two of the three speakers. Furthermore, as it is clear that their combination improves the detection performance, only the combination of the two (i.e. Comp. 1 & 2) will be considered for the remainder of the study.

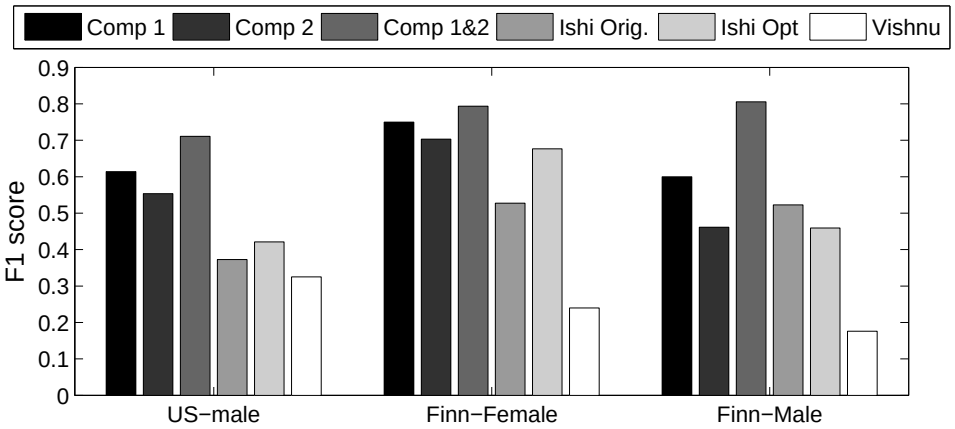


Figure 8.8: F1 scores for the six different detection methods on the TTS database.

Ishi's algorithm shows strong performance on the Finnish female speaker, with an improvement over the original algorithm when the parameters are used as inputs to the decision tree classifier. A small improvement is also seen for the American male speaker and a slight reduction for the Finnish male.

The algorithm described in Vishnubhotla and Espy-Wilson (2006) (labelled Vishnu.) displayed relatively low performance compared to the other algorithms, with a consistently low F1 score. This was found to be largely due to an high number of false alarms. In the original study (Vishnubhotla and Espy-Wilson, 2006) the authors use the terms 'irregular phonation' and creak interchangeably, but upon examination of the false alarms it was found that a broader class of 'irregular phonation' types were detected, many of which did not match the auditory criterion used in this study and in Ishi et al. (2008b). As a result this algorithm was excluded from the remainder of the analysis as the grounds for

comparison seemed tenuous.

8.6.2 Effect of post-processing

The effect of the post-processing step on the event level metrics, summed across all speakers, on the three detection methods is shown in Table 8.2. Although an increase in misses can be observed, there is, nevertheless, a substantial reduction (in the region of 50 %) in false alarms for the three methods. It is clear that the use of the decision tree classifier with the parameters from Ishi et al. (2008b) as input features (i.e. Ishi Opt.) causes both an increase in hits but also a substantial increase in false alarms. The post-processing step considerably reduces the number of false alarms, although Ishi Opt. still displays more false alarms than the other two methods. As will be seen in Section 8.6.3 a large proportion of these false alarms come from the Finnish male speaker in the TTS database, where Ishi Opt. frequently produces misdetections in low pitch, non-creak, voiced segments.

Table 8.2: Hits, misses and false alarms totalled across all speakers for the four detection methods. Results are shown without and with additional post-processing.

	WITHOUT POST-PROCESSING			WITH POST-PROCESSING		
Method	**Hits**	**misses**	**false alarms**	**Hits**	**misses**	**false alarms**
Comp. 1&2	2320	426	2039	2221	525	1009
Ishi Orig.	1808	938	2311	1617	1129	1206
Ishi Opt.	2264	482	7142	2086	660	3561

A slight improvement of F1 score has been noticed for all techniques. The F1 score for Comp. 1 & 2 has been improved on average by 0.015 $\pm$ 0.007 (standard deviation), for Ishi Orig. by 0.001 $\pm$ 0.011, and for Ishi Opt by 0.019 $\pm$ 0.016. There is a rather minor effect of the post-processing on F1, although there is a considerable improvement at the event level. This can be explained by the fact that the majority of the false alarms shown in the event level results were short in duration and, hence, their removal did not contribute strongly to the resulting F1 scores.

As the post-processing brought an improvement to the three methods (particularly in terms of event level metrics) it will be used in with these three methods (i.e. Comp. 1 & 2, Ishi Orig. and Ishi Opt.) for remainder of the study.

8.6.3 Detailed survey of detection performance

The frame level results for the three detection methods are presented in Figure 8.9, with the event level results shown in Table 8.3. Note that the F1 score, shown in the right column of Figure 8.9, gives the clearest impression of the performance of the detection methods in a single measure and, hence, will receive most attention.

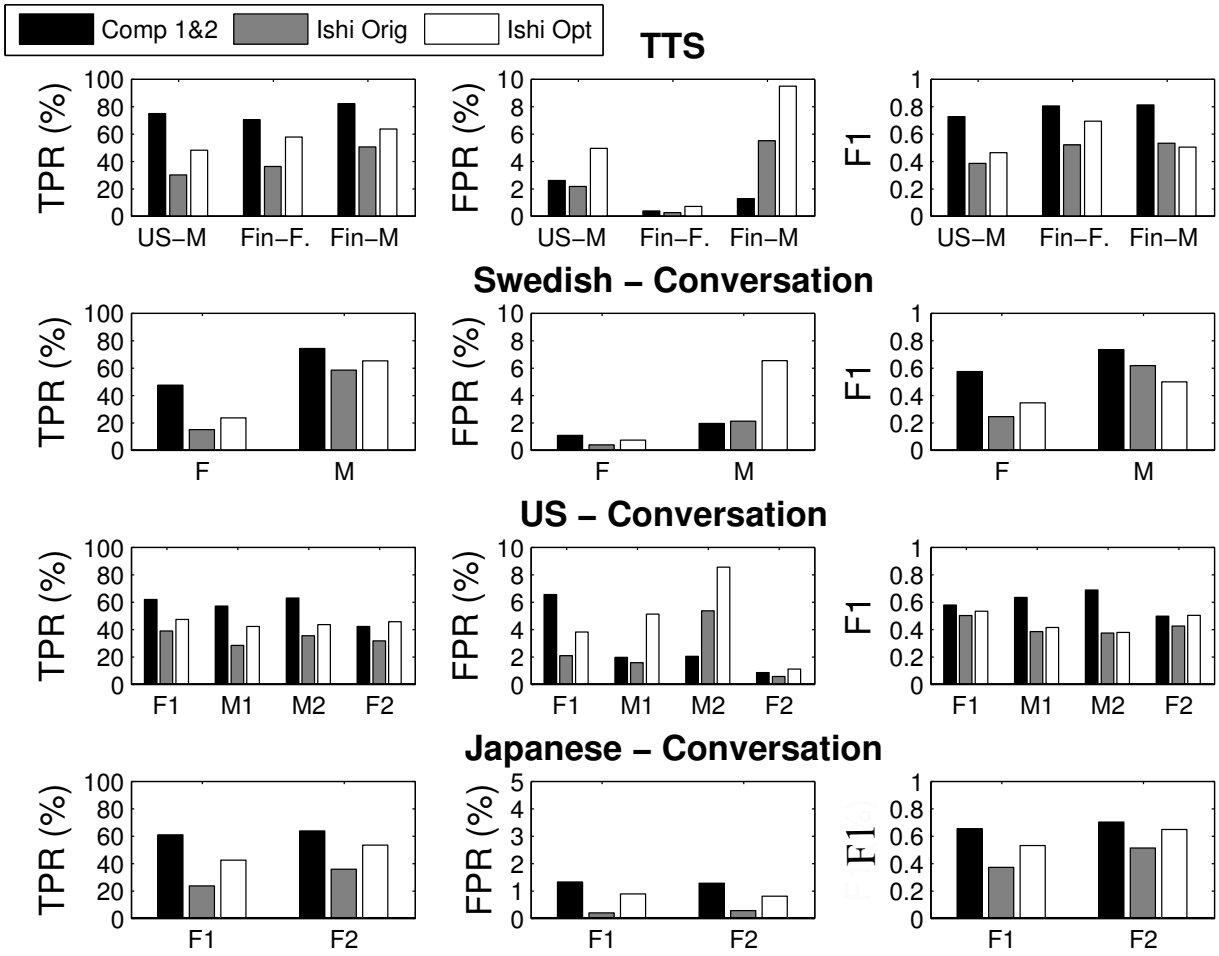


Figure 8.9: Frame level metrics (i.e. TPR, FPR and F1) for the three detection methods shown for each speaker in the four databases. M is used for male and F for female.

Considering Figure 8.9 one can observe that the proposed detection method (Comp. 1 & 2) produced higher F1 scores than the two comparison methods for every speaker, with the exception of Female 2 from the US database where Comp. 1 & 2 and Ishi Opt. produced the same F1 score of 0.49). This is due to a comparatively high true positive rate (TPR) and low false positive rate (FPR). However, for female 1 in the US database there is relatively high FPR for Comp. 1 & 2. This was investigated and it was found

that a large proportion of the false alarms contained noises from the speaker's mouth colliding with the microphone or other background noises, mostly occurring as the person was speaking. Additional features for detecting such noises would help reduce the number of false alarms of this kind.

This general trend was also supported in the event level results (Table 8.3) where Comp. 1 & 2 gave a higher number of hits and a lower number of misses than Ishi Orig. for every speaker. This was often combined with a lower number of false alarms. Compared with Ishi Opt., Comp. 1 & 2 had a more similar level of hits and misses, although it generally led to a much lower number of false alarms.

The F1 score for Ishi Opt was higher than Ishi Orig for almost every speaker (the exception being the Finnish male speaker from the TTS database). This was due to an increased TPR over Ishi Orig, but this was also coupled with an increased FPR for every speaker.

Table 8.3: Event level results (i.e. hits, misses and false alarms) for the three creaky voice detection methods for each speaker in the four databases.

		COMP 1& 2			ISHI ORIG			ISHI OPT		
Database	**Speaker**	**hits**	**misses**	**FAs**	**hits**	**misses**	**FAs**	**hits**	**misses**	**FAs**
TTS	US-M	153	12	42	111	54	97	145	20	312
	Finn-F	211	51	7	171	91	13	220	42	90
	Finn-M	193	20	53	173	40	477	195	18	1098
Swedish	F	378	142	192	174	346	72	247	273	225
	M	237	33	103	221	49	157	240	30	652
US	F1	192	37	212	154	75	84	178	51	233
	M1	75	12	53	55	32	60	73	14	257
	M2	89	16	35	64	41	182	71	34	294
	F2	37	17	38	31	23	18	39	15	68
Japanese	F1	413	132	167	274	271	28	422	123	250
	F2	243	53	107	189	107	18	256	40	82

For Ishi Opt. there was generally a higher FPR and number of false alarms for male speakers. This was particularly true for the Finnish male speaker in the TTS database. This speaker had the lowest pitch of all the speakers with a mean f_0 typically around 80 Hz. The false alarms were investigated and they were found to be frequently due to a lack of distinction between low pitch, voiced, non-creaky segments and true creaky segments. The IFP parameter, which is utilised in these methods for differentiating normal voiced regions from creaky regions, frequently produced very low values in these low pitch regions

which led to these false alarms. Further features to help disambiguate these two classes would certainly improve the detection performance when using the parameters from Ishi et al. (2008b) for detecting creak.

In order to investigate whether the F1 scores for the proposed detection method were significantly higher than those from the two comparison methods, A one-way ANOVA was carried out with F1 score treated as the dependent variable and detection method as the independent variable. This revealed that the detection method had a significant effect on the F1 score [$F_{(2,30)} = 15.002, p < 0.001$], and subsequent pairwise comparisons carried out using Tukey's Honestly Significant Difference (HSD) test revealed that the Comp. 1 & 2 method gave significantly higher F1 scores than both Ishi Orig. ($p < 0.001$) and Ishi Opt. ($p < 0.01$).

Also, to investigate whether gender had a significant effect on the F1 scores for each of the methods, a two-way ANOVA was carried out with F1 score as the dependent variable and with detection method and gender as the independent variables. Although lower mean F1 scores were observed for females with the Comp. 1 & 2 method, and lower means for males in the two Ishi methods, the two-way ANOVA revealed no significant effect of gender [$F_{(1,27)} = 0.057, p = 0.81$]. However, repeating the same test with FPR as the dependent variable gender was found to have a significant effect [$F_{(1,27)} = 25.078, p < 0.001$] and pairwise comparisons (using Tukey's HSD) revealed that for Comp. 1& 2 and Ishi Opt. FPR was significantly higher ($p < 0.05$) for males.

8.7 Results on degraded data

The effect of additive noise on the detection performance of the three methods is illustrated in Figure 8.10. It can be observed that in the white noise condition the Comp. 1 & 2 method achieves the highest F1 score at all levels of signal to noise ratio (SNR). Ishi Opt. achieves a slightly higher F1 than Ishi Orig. down to an SNR of 10 dB. For the babble noise condition again the Comp. 1 & 2 method attains a higher F1 than the two comparison methods down to an SNR of 20 dB. However, at SNR of 10 dB the F1 for Comp. 1 & 2 falls slightly below that of Ishi Opt. All methods deteriorate severely at 0 dB SNR.

The findings here are encouraging for the proposed method as they suggest Comp. 1 & 2 can provide superior detection of creaky voice even in conditions with moderately high levels of noise. Babble noise is shown to have a stronger negative effect on the performance of the three methods compared to white noise. This is likely due to the more pronounced low frequency characteristic of babble compared to white noise which more severely affects

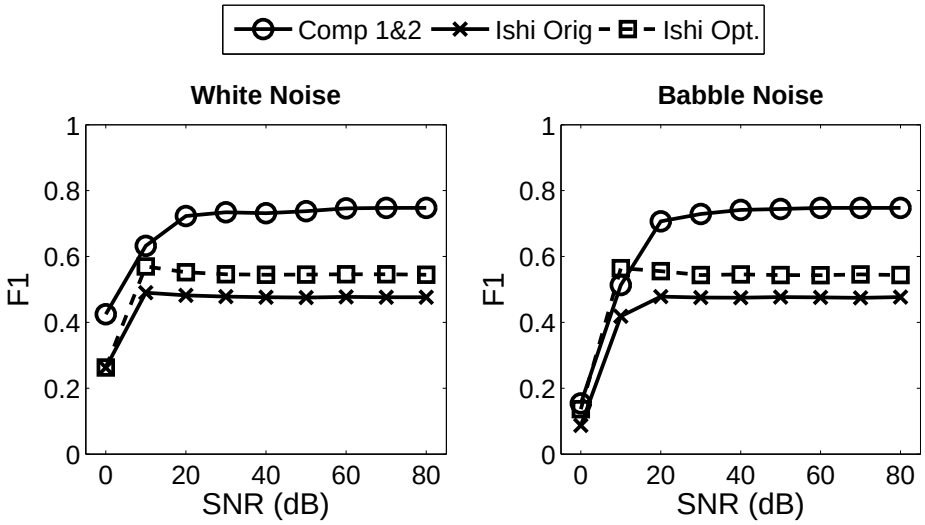


Figure 8.10: Effect of white noise (left panel) and babble noise (right panel) on the F1 score (averaged across the three speakers in the TTS database) achieved by the three creaky voice detection methods.

the parameters used in the three methods at low SNR levels.

8.8 Discussion and conclusion

This chapter presents a new method for automatically detecting creaky voice in speech signals by exploiting characteristics of the LP-residual signal, namely the presence of secondary peaks and long glottal pulses with prominent impulse-like excitation peaks. Resonators were applied to the LP-residual and two parameters were derived from the characteristics of the resonator output. These parameters were then used as input features to a decision tree classifier. The resulting detection performance was shown to significantly outperform existing creaky voice detection methods on a large range of speech data covering different speakers, gender, languages, recording conditions and speaking styles (i.e. read vs conversational speech). These findings build on the initial promising results reported in Drugman et al. (2012a).

Furthermore, the new method demonstrated robustness to white noise, with the highest performance across all SNR levels, and to babble noise, with improved detection over the comparison methods down to 20 dB SNR.

The inclusion of the parameters derived using the methods described in Ishi et al. (2008b) in a decision tree classifier which was optimised on training sets brought some improvement to the overall detection performance compared to the original algorithm, with the original threshold settings. However, despite this improvement there was still an

increase in the level of false alarms. The inclusion of further features, in such a classifier, which could help disambiguate non-creaky voiced segments and creaky segments would certainly bring a further improvement to the detection performance.

8.9 Applications

The new creaky voice detection method described here has a very strong potential for including creaky voice in speech technology applications. My initial collaborative work on modelling the creaky excitation for statistical parametric speech synthesis involved the use of manual creak annotation. By apply the new creaky voice detection method this process could be automated. My on-going collaborative work on this topic has involved investigate the extent to which contextual factors (phoneme, word stress, position in the sentence, prosodic context, etc) can be used to predict locations of creak. This has also involved the use of the detection algorithm. Finally, it is hoped that the new algorithm can be used to help quantitatively study the use of creaky voice on larger volumes of data and to investigate its potential in applications like speaker identification.

Relevant publications

- Kane, J., Drugman, T., Gobl, C., (2013) Improved automatic detection of creak, *Computer Speech and Language* 27(4), pp. 1028-1047.
- Drugman, T., Kane, J., Gobl, C., (2012) Resonator based creaky voice detection, *Proceedings of Interspeech, Portland, Oregon, USA.*
- Drugman, T., Kane, J., Gobl, C., (2012) Modeling the creaky excitation for parametric speech synthesis, *Proceedings of Interspeech, Portland, Oregon, USA.*
- Kane, J., Pápay, K., Hunyadi, L., Gobl, C., (2011) On the use of creak in Hungarian spontaneous speech, *Proceedings of ICPhS, Hong Kong.*

Part IV

Conclusions

Chapter 9

General discussion and conclusions

This thesis documents a series of developments in glottal source and voice quality analysis. Specifically, the objective was to describe and evaluate novel algorithms to improve automatic analysis of the voice.

Initially descriptions were provided of the physiological and acoustic correlates of different vocal settings (Chapter 2). Chapter 2 also summarises the source-filter theory (Fant, 1960) on which much of the glottal source and voice quality analysis in this thesis (and indeed elsewhere in the literature) is based. Then in Chapter 3 the importance of glottal source and voice quality variation in spoken communication is emphasised. It is also illustrated here how characterisation of variations in vocal settings can be used in speech technology and, further, how more robust algorithms would facilitate fuller exploitation of these important aspects of speech.

The subsequent chapters then proceed to describe and evaluate novel algorithms to help improve the robustness of glottal source and voice quality analysis. The detection of glottal closure instants (GCIs), which is a fundamental starting point for glottal synchronous analysis, is the focus of Chapter 4. A new algorithm, SE-VQ, is proposed, designed to handle the different acoustic characteristics of a range of phonation types. The method was shown to have similar performance to the state-of-the-art on standard speech databases and to produce an important reduction in false alarms in creaky voice regions.

Some improvement in the precision of GCI localisation was also observed. However a higher accuracy error still occurs for certain phonation types, e.g., breathy and harsh voice. Note that this was also true for other state-of-the-art algorithms. The dynamic programming component of the algorithm appears to be a useful tool for maintaining consistent positioning of the GCIs, although a reasonable starting point (i.e. a prominent LP-residual peak) is required in order to obtain this consistent performance. The problem

of improving GCI detection for speech where there are no clear residual peaks remains an open and important question for future research.

The SE-VQ algorithm was subsequently used for locating GCIs in a new method for parameterisation of an estimated glottal source derivative using the LF model (DyProg-LF, Chapter 5). The new algorithm was shown to provide consistent glottal source parameterisation on a carefully controlled dataset, with reference values obtained from manual analysis, as well as on a larger dataset where reference values were automatically derived from EGG signals. In particular the dynamic programming component appeared to be a suitable mechanism for maintaining a consistency in the parameter contour in regions of high stationarity. This was a problem noted in previous LF model fitting algorithms. An area yet to be examined is the design and evaluation of glottal modelling methods for more irregular phonation types such as creaky or harsh voice. There may also be scope for extending this algorithm to be used in the characterisation of certain voice pathologies. Such research would look to extend some recent methodological developments for the estimation and characterisation of the glottal source for dysphonic speakers (see e.g., Dubuisson, 2012);

Note that the same dynamic programming algorithm was exploited for both the SE-VQ and DyProg-LF methods. The work by Talkin (1995) was the inspiration for using this particular dynamic programming algorithm. Even though the f_0 and formant trackers were developed by Talkin in the 1980s and 1990s, they are still widely used today. Although this dynamic programming algorithm was mainly used by Talkin to avoid gross errors (i.e. octave doubling/halfing in f_0 tracking and the selection of the number of formants in the formant tracking) evidence in the present thesis suggests that by setting appropriate target and transition costs, and their respective weights, this algorithm can be used to avoid more subtle errors.

This parameterisation algorithm, as well as other methods for glottal source parameterisation and glottal inverse filtering were evaluated using a range of different experiments in Chapter 6. Model fitting methods, and in particular the newly proposed method, were shown to be considerably more effective at discriminating breathy to tense voice than has previously been reported. Furthermore, the DyProg-LF method was shown to provide more natural resynthesised utterances than a comparison method.

Chapter 6 also highlights the various shortcomings of the three main glottal inverse filtering methods. It is apparent that the inverse filtering problem is yet to be sufficiently solved. This is despite many years of research attention to this topic. Nevertheless, by improving the robustness of subsequent parameterisation approaches, moderate errors in the inverse filtering may be negated somewhat. The development of algorithms for both

glottal inverse filtering and glottal source parameterisation is still hampered by the lack of robust, automatically derived reference values. The approach in this thesis was to evaluate from several different viewpoints in order to achieve a rounded picture of the performance of the algorithms. Even still, research in this area would certainly benefit from a concerted effort to find robust evaluation metrics and procedures.

In Chapter 7, a new parameter, the Maxima Dispersion Quotient (MDQ), is described for differentiating breathy to tense voice using features of a wavelet transform. MDQ was shown to be more effective than existing parameters for discriminating the voice qualities and, in particular, on continuous speech. The effectiveness of the comparison parameters degraded severely for continuous speech. These methods require prior glottal inverse filtering and as a result the findings here also point the shortcomings of automatic glottal inverse filtering of running speech.

The final experimental chapter, Chapter 8, describes two acoustic parameters derived from the Linear Prediction (LP) residual for characterising creaky voice. These parameters were used as inputs to a decision tree classifier which was shown to significantly improve classification of creak, over the state-of-the-art, on a range of speech data varying in terms of speaker, gender, language, recording condition and speaking style. The results in particular showed an increase in true positives for the proposed method. A variety of acoustic characteristics can give rise to the perception of creak and, hence, these findings suggest that the proposed parameters cover much of this variety. Even still, the percentage of true positives can be increased further and one obvious direction would be to include parameters from comparison methods as part of a larger feature vector to be used in a classifier. Further gains may be achieved by optimising the classifier type and settings. Nevertheless, the approach used in this thesis involving simply two acoustic parameters and a straightforward classification method already achieved good results.

One of the main aims of the thesis was to develop new methods for detecting various commonly occurring voice qualities and the new algorithms presented in Chapter 7 and 8 go a long way to achieving this aim. Despite this, the detection of certain voice qualities (e.g., whisper, harsh voice) has not been covered in this thesis. Some recent developments in the literature have looked to tackle this problem of detecting other voice qualities (see for example Ishi et al., 2011; Obin, 2012), but there still remains much scope to improve the performance of these methods to help further facilitate the use of voice quality in speech technology.

A final contribution of the thesis is a software package, the *Voice analysis toolkit*, which has been made publicly available. The toolkit contains the novel algorithms described in Chapters 4 - 8 and is intended to encourage usage in applied studies, experimentation and

feedback. The README file contained in the toolkit is also given in Appendix B. There are, of course several other toolkits for carrying voice analysis available. For instance, the VOICEBOX[1] toolkit which provides a large set of functions for carrying out speech processing, including LPC, f_0 tracking and peak-picking algorithms, the APARAT[2] (Airas, 2008), SKY[3] (Kreiman et al., 2006) and Voice Sauce [4] toolkits which provide graphical user interfaces (GUIs) for carrying out glottal source analysis and a recent toolkit, GLOAT[5], which contains a range of novel algorithms, including: f_0 tracking, voicing decision, GCI detection, signal polarity detection, complex-cepstrum based decomposition, etc.

The *Voice analysis toolkit* produced in the present research is complementary to these existing toolkits. In fact both GLOAT and VOICEBOX are utilised in the current toolkit. Furthermore, the novel algorithms contained in this new toolkit can be easily integrated into existing GUIs. These methods have also been integrated into a new GUI for glottal source analysis, which allows both manual and automatic approaches, that has been developed at the Phonetics and Speech Laboratory in Trinity College.

9.1 Future directions

From the above discussion, it may be clear that despite the progress made as part of this thesis there remains a host of research problems outstanding, in terms of the development of methods for glottal source and voice quality analysis. For glottal source analysis, a major outlying problem remains to be the effectiveness of glottal inverse filtering of continuous speech. In terms of parameterisation using glottal models an important challenge is to design methods appropriate for the modelling of irregular (in terms of periodicity and amplitude modulation) phonation types (e.g., harsh and creaky voice). For the detection of changes in voice quality, the design of further features which are suitable for characterising speech recorded in less than ideal conditions is a clear research goal. Furthermore, such new features should extend to other non-modal voice qualities. This would benefit the inclusion of voice quality in speech technology applications.

In terms of application of the methods developed within this thesis, on-going and future research will involve exploiting the newly developed algorithms for different purposes. An illustration was given in Chapter 1 of how fine-grained glottal synchronous methods (Chapters 4 - 6) could be combined with coarse-grained voice quality detection methods

[1] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[2] http://sourceforge.net/projects/aparat/
[3] http://www.surgery.medsch.ucla.edu/glottalaffairs/software.htm
[4] http://www.ee.ucla.edu/~spapl/voicesauce/
[5] http://tcts.fpms.ac.be/~drugman/Toolbox/

(Chapters 7 - 8) for the purpose of improving statistical parametric speech synthesis. In fact the author's on-going research has exploited methods from both of these parts of the thesis for exactly this purpose.

The study described in Drugman et al. (2012b) proposes a method for incorporating creaky voice in parametric speech synthesis. Initially creaky regions need to be identified (which can be done using the method described in Chapter 8). As GCIs are required for this method, the SE-VQ algorithm (Chapter 4) is also exploited in this study. Indeed many other potential speech technology applications can be envisaged which would draw on findings from both these parts of the thesis.

Other future directions will involve the use of the SE-VQ algorithm (Chapter 4) and the DyProg-LF parameterisation method (Chapter 5) to help automate and semi-automate the measurement workflows carried out in linguistic studies on the role of the glottal source in the prosody of spoken language, building on previous studies (Gobl, 1988; Yanushevskaya et al., 2009, 2010; Ní Chasaide et al., 2011). As was stated above, these methods have been integrated into the in-house graphical user interface (GUI) at the Phonetics and Speech Laboratory for this very purpose.

Furthermore, it is envisaged that the SE-VQ and DyProg-LF methods will be exploited in parametric speech synthesis and voice modification systems. It is hoped that these methods will help provide more natural rendering of the speakers voice quality variation and will facilitate the design of platforms that allow flexible manipulation of voice characteristics at synthesis time.

Some previous work in collaboration with other researchers involved the clustering and separation of speaking styles using voice quality features (Székely et al., 2012b; Scherer et al., 2012b). The hope is to continue these collaborations and to utilise the methods for discriminating breathy, tense and creaky voice (Chapters 7 - 8) to help improve this research.

Finally, it has been mentioned in this thesis that the study of the usage of voice quality has been hampered by the lack of robust detection algorithms. It is hoped that the methods described in Chapters 7 - 8 can be leveraged, both by the present author and by other people in the research community, to help carry out more quantitative analysis on the use of voice quality variation in larger volumes of data. The new *Voice analysis toolkit* should help facilitate this process.

Bibliography

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

Agiomyrgiannakis, Y. and Rosec, O. ARX-LF-based source-filter methods for voice modification and transformation. *Proceedings of ICASSP, Taipei, Taiwan*, pages 3589–3592, 2009.

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

Airas, M. and Alku, P. Comparison of multiple voice source parameters in different phonation types. *Proceedings of Interspeech 2007, Antwerp, Belgium*, pages 1410–1413, 2007.

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

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

Alku, P. and Vilkman, E. Estimation of the glottal pulseform based on discrete all-pole modeling. *Proceedings of the Third International Conference on Spoken Language Processing*, pages 1619–1622, 1994.

Alku, P., Strik, H., and Vilkman, E. A new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79, 1997.

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

Alku, P., Magi, C., Yrttiaho, S., Bäckström, T., and Story, B. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009.

Allen, D. R. and Strong, W. J. A model for the synthesis of natural sounding vowels. *Journal of the Acoustical Society of America*, 78:58–69, 1985.

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

Ananthapadmanabha, T. V. Acoustic analysis of voice source dynamics. *KTH, Speech Transmission Laboratory, Quarterly Report*, 25(2-3):1–24, 1984.

Baayen, R. H. *Analyzing Linguistic Data. A Practical Introduction to Statistics*. Cambridge University Press, 2008.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.

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

Böhm, T. and Shattuck-Hufnagel, S. Listeners recognize speakers' habitual utterance-final voice quality. *Proceedings of ParaLing07*, pages 29–34, 2007.

Böhm, T., Both, Z., and Németh, G. Automatic classification of regular vs. irregular phonation types. *Advances in nonlinear speech processing*, pages 43–50, 2010.

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

Brackett, I. *The growth of inflammation of the vocal folds accompanying easy and harsh production of the voice*. M. A. Thesis, Northwestern University, 1940.

Braunschweiler, N. and Buchholz, S. Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality. *Proceedings of Interspeech, Florence, Italy*, pages 1821–1824, 2011.

Breiman, L., Stone, C. J., Olshen, R. A., and Friedman, J. H. *Classification and Regression Trees*. Wadsworth Inc., 1984.

Brent, R. P. *Algorithms for Minimization Without Derivatives.* Englewood Cliffs, NJ: Prentice-Hall, 1973.

Brown, P. and Levinson, S. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

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

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

Cabral, J.P., Renals, S., Yamagishi, J., and Richmond, K. HMM-based speech synthesiser using the LF-model of the glottal source. *Proceedings of ICASSP, Prague, Czech Republic*, pages 4704–4707, 2011b.

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

Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I., and Lin, Q. Voice source rules for text-to-speech synthesis. *Proceedings of the IEEE International conference on acoustic speech signal processing, Glasgow, Scotland*, 1:223–223, 1989.

Carlson, R., Gustafson, K., and Strangert, E. Cues for hesitation in speech synthesis. *Proceedings of Interspeech, Pittsburgh, USA*, pages 1300–1303, 2006.

Catford, J. *Phonation types: the classification of some laryngeal components of speech production*, pages 26–37. In honour of Daniel Jones. London: Longmans, 1964.

Chen, G., Kreiman, J., Shue., Y. L., and Alwan, A. Acoustic correlates of glottal gaps. *Proceedings of Interspeech, Florence, Italy*, pages 2673–2676, 2011.

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

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

Cruttenden, A. *Intonation*. Cambridge University Press, 1986.

d'Alessandro, C. and Sturmel, N. Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude. *Sadhana*, 36(5):601–622, 2011.

Degottex, G. *Glottal source and vocal tract separation - Estimation of glottal parameters, voice transformation and synthesis using a glottal model*. PhD thesis, IRCAM, Paris, 2010.

Degottex, G., Roebel, A., and Rodet, X. Shape parameter estimate for a glottal model without time position. *SPECOM*, pages 345–349, 2009.

Degottex, G., Roebel, A., and Rodet, X. Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter. *Proceedings of ICASSP, Prague, Czech Republic*, pages 5128–5131, 2011a.

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

Degottex, G., Lanchantin, A., Roebel, A., and Rodet, X. Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis. *Speech Communication*, 55(2):278–294, 2012.

del Pozo, A. *Voice source and duration modelling for voice conversion and speech repair*. PhD thesis, Cambridge University, 2008.

Deshmukh, D., Espy-Wilson, C., Salomon, A., and Singh, J. Use of temporal information: Detection of periodicity, aperiodicity and pitch in speech. *IEEE Transactions on Audio Speech and Language processing*, 13(5):776–786, 2005.

Doval, B., d'Alessandro, C., and Henrich, N. The voice source as a causal/anticausal linear filter. *VOQUAL '03, Geneva, Switzerland*, pages 16–20, 2003.

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

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

Drugman, T. and Dutoit, T. On the potential of glottal signatures for speaker recognition. *Proceedings of Interspeech, Makuhari, Japan*, pages 2106–2109, 2010.

Drugman, T. and Dutoit, T. Oscillating statistical moments for speech polarity detection. *Proceedings of Non-Linear Speech Processing Workshop (NOLISP11), Las Palmas, Gran Canaria, Spain*, pages 48–54, 2011.

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

Drugman, T., Dubuisson, T., and Dutoit, T. On the mutual information between source and filter contributions for voice pathology detection. *Proceedings of Interspeech, Brighton, UK*, pages 1463–1466, 2009b.

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

Drugman, T., Bozkurt, B., and Dutoit, T. A comparative study of glottal source estimation techniques. *Computer Speech and Language*, 26:20–34, 2011.

Drugman, T., Kane, J., and Gobl, C. Resonator-based creaky voice detection. *Proceedings of Interspeech, Portland, Oregon, USA*, 2012a.

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

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

Dubuisson, T. *Glottal source estimation and automatic detection of dysphonic speakers*. PhD thesis, University of Mons, Belgium, 2012.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., and House, D. Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. *Proceedings of LREC, Malta*, pages 2992–2995, 2010.

Edmondson, J.A. and Esling, J.H. The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies. *Phonology*, 23(2)(2):157–191, 2006.

El-Jaroudi, A. and Makhoul, J. Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 32(2):411–423, 1991.

Elliot, J. R. The application of a Bayesian approach to auditory analysis in forensic speaker identification. *Proceedings of the 9th Australian International Conference on Speech Science and Technology*, pages 315–320, 2002.

Esling, J.H. Laryngographic study of phonation type and laryngeal configuration. *Journal of the International Phonetic Association*, 14:56–73, 1984.

Esling, J.H. and Harris, J.G. An expanded taxonomy of states of the glottis. *Proceedings of ICPhS, Barcelona, Spain*, 1:1049–1052, 2003.

Espy-Wilson, C., Manocha, S., and Vishnubhotla, S. A new set of features for text-independent speaker identification. *Proceedings of Interspeech (ICSLP), Pittsburgh, Pennsylvania, USA*, pages 1475–1478, 2006.

Fairbanks, G. *Voice and Articulation Drill Book (2nd edition)*. Harper and Row, New York, 1960.

Fan, X. and Hansen, J. Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams. *Speech communication*, 55(1):119–134, 2013.

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

Fant, G. Glottal source and excitation analysis. *KTH, Speech Transmission Laboratory, Quarterly Report*, 20(1):85–107, 1979.

Fant, G. and Lin, Q. Glottal source - vocal tract acoustic interaction. *KTH, Speech Transmission Laboratory, Quarterly Report*, 28(1):13–27, 1987.

Fant, G., Liljencrants, J., and Lin, Q. A four parameter model of glottal flow. *KTH, Speech Transmission Laboratory, Quarterly Report*, 4:1–13, 1985a.

Fant, G., Lin, Q., and Gobl, C. Notes on glottal flow interaction. *KTH, Speech Transmission Laboratory, Quarterly Report*, 2-3:21–45, 1985b.

Fant, G., Kruckenberg, A., Liljencrants, J., and Båvegård, M. Voice source parameters in continuous speech. transformation of LF-parameters. *Proc of ICSLP-94, International conference on spoken language processing Yokohama, Japan*, 3:1451–1454, 1994.

Fant, G., Liljencrants, J., and Lin, Q. The LF-model revisited. transformations and frequency domain analysis. *KTH, Speech Transmission Laboratory, Quarterly Report*, 2-3:119–156, 1995.

Félix Torres, J. and Moore, E. Speaker discrimination ability of glottal waveform features. *Proceedings of Interspeech, Portland, Oregon, USA*, 2012.

Flanagan, J. L. *Speech Analysis, Synthesis and Perception.* Springer-Verlag, 2nd edition, 1972.

Fønagy, I. Mimik auf glottaler ebene. *Phonetica*, 8:209–219, 1962.

Fujimoto, M. and Maekawa, K. Variation of phonation types due to paralinguistic information: an analysis of high speed video images. *Proceedings of ICPhS, Barcelona, Spain*, 2003.

Fujisaki, H. and Ljungqvist, M. Proposal and evaluation of models of the glottal source waveform. *Proceedings of the IEEE International conference on acoustic speech signal processing, Tokyo, Japan*, 4:1605–1608, 1986.

Gerratt, B. R. and Kreiman, J. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, 29:365–381, 2001.

Ghosh, P.K., Tsiartas, A., and Narayanan, S. Robust voice activity detection using long-term signal variability. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):600–613, 2011.

Gobl, C. Voice source dynamics in connected speech. *KTH, Speech Transmission Laboratory, Quarterly Report*, 29(1):123–159, 1988.

Gobl, C. A preliminary study of acoustic voice quality correlates. *KTH, Speech Transmission Laboratory, Quarterly Report*, 4:9–21, 1989.

Gobl, C. The voice source in speech communication. *Ph. D. Thesis, KTH Speech Music and Hearing, Stockholm*, 2003.

Gobl, C. Modelling aspiration noise during phonation using the LF voice source model. *Proceedings of Interspeech, Pittsburgh, USA*, pages 965–968, 2006.

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

Gobl, C. and Ní Chasaide, A. Techniques for analysing the voice source. In Hardcastle, W. J. and Hewlett, N., editors, *Coarticulation: Theory, data and techniques*, chapter 15, pages 300–320. Cambridge University Press, 1999a.

Gobl, C. and Ní Chasaide, A. Perceptual correlates of source parameters in breathy voice. *Proceedings of ICPhS, San Francisco*, pages 2437–2440, 1999b.

Gobl, C. and Ní Chasaide, A. Amplitude-based source parameters for measuring voice quality. *VOQUAL, Geneva, Switzerland*, pages 151–156, 2003a.

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

Gobl, C. and Ní Chasaide, A. Voice source variation and its communicative functions. In Hardcastle, W. J., Laver, J., and Gibbon, F. E., editors, *The handbook of phonetic sciences (2nd Edition)*, pages 378–423. Oxford, Backwell, 2010.

Gordon, M. and Ladefoged, P. Phonation types: a cross-linguistic review. *Journal of Phonetics*, (29):383–406, 2001.

Guruprasad, S., Yegnanarayana, B., and Murty, K.S.R. Detection of instants of glottal closure using characteristics of excitation source. *Proceedings of Interspeech, Antwerp, Belgium*, pages 554–58, 2007.

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

Hanson, H., Stevens, K., Kuo, H., Chen, M., and Slifka, J. Towards models of phonation. *Journal of Phonetics*, (29):451–480, 2001.

Hanson, H. M. Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, 10(1):466–481, 1997.

Hardcastle, W. J. *Physiology of Speech Production*. Academic Press Inc., London, 1976.

Heldner, M., Edlund, J., Hjalmarsson, A., and Laskowski, K. Very short utterances and timing in turn-taking. *Proceedings of Interspeech, Florence, Italy*, pages 2837–2840, 2011.

Henrich, N., d'Alessandro, C., and Doval, B. Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. *Proceedings of EUROSPEECH, Scandanavia*, 2001.

Henrich, N., d'Alessandro, C., Doval, B., and Castellengo, M. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of the Acoustical Society of America*, 115(3):1321–1332, 2004.

Hoit, J. and Hixon, T. Body type and speech breathing. *Journal of Speech and Hearing Research*, 29:313–324, 1986.

Hollien, H. and Wendahl, R. W. Perceptual study of vocal fry. *Journal of the Acoustical Society of America*, 47(3):506–509, 1968.

Holmes, J. N. Formant synthesizers: Cascade or parallel? *Speech Communication*, 2(4), 251-273 1983.

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

Hudgins, C. V. and Stetson, R. H. Relative speed of articulatory movements. *Archives néerlandaises de phonétique experimentale*, 13:85–94, 1937.

Hui-Ling, Lu. *Toward a high quality singing synthesizer with vocal texture control*. PhD thesis, Stanford University, 2002.

Ishi, C. T. Analysis of autocorrelation-based parameters for creaky voice detection. *Proceedings of Speech Prosody, Nara, Japan*, pages 643–646, 2004.

Ishi, C. T., Ishiguro, H., and Hagita, N. Proposal of acoustic measures for automatic detection of vocal fry. *Proceedings of Interspeech, Lisbon, Portugal*, pages 481–484, 2005.

Ishi, C. T., Ishiguro, H., and Hagita, N. Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality. *Speech communication*, 50(6):531–543, 2008a.

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

Ishi, C. T., Ishiguro, H., and Hagita, N. Improved acoustic characterization of breathy and whispery voices. *Proceedings of Interspeech, Florence, Italy*, pages 2965–2968, 2011.

Ishi, C.T., Sakakibara, K.I., Ishiguro, H., and Hagita, N. A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):47–56, 2008b.

Itakura, F. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23:67–72, 1975.

Itakura, F. and Saito, S. Analysis synthesis telephony based on the maximum likelihood method. *Proceedings of the International Congress on Acoustics, Tokyo, Japan*, 1968.

Ito, M. Politeness and voice quality - the alternative method to measure aspiration noise. *Proceedings of Speech Prosody, Nara, Japan*, 2004.

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

Kane, J. and Gobl, C. Automatic parameterisation of the glottal waveform combining time and frequency domain measures. *Proceedings of Maveba 2009, Florence*, pages 91–94, 2009.

Kane, J. and Gobl, C. Identifying regions of non-modal phonation using features of the wavelet transform. *Proceedings of Interspeech, Florence, Italy*, pages 177–180, 2011.

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

Kane, J., Kane, M., and Gobl, C. A spectral LF model based approach to voice source parameterisation. *Proceedings of Interspeech, Makuhari, Japan*, pages 2606–2609, 2010.

Kane, J., Yanushevskaya, I., Ní Chasaide, A., and Gobl, C. Exploiting time and frequency domain measures for precise voice source parameterisation. *Proceedings of Speech Prosody, Shanghai, China*, 1:143–146, 2012.

Kasuya, H., Yoshizawa, M., and Maekawa, K. Roles of voice source dynamics as a conveyer of paralinguistics features. *Proceedings of Spoken Language Processing (ICSLP '00)*, 2000.

Kawahara, H. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. *Proceedings of ICASSP*, 2:1303–1306, 1997.

Kay, S.M. *Modern Spectral Estimation: Theory and Application*. Prentice Hall Englewood Cliffs, NJ, 1988.

Klatt, D. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:13–33, 1980.

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

Kominek, J. and Black, A. The CMU ARCTIC speech synthesis databases. *ISCA speech synthesis workshop, Pittsburgh, PA*, pages 223–224, 2004. URL `http://festvox.org/cmuarctic/`.

Konvalinka, I. S. and Mataušek, M. R. Simultaneous estimation of poles and zeros in speech analysis and ITIF-iterative inverse filtering algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-27(5):485–492, 1979.

Kounoudes, A., Naylor, P., and Brookes, M. The DYPSA algorithm for estimation of glottal closure instants i n voiced speech. *Proceedings of ICASSP, Orlando, Florida*, 11: 349–352, 2002.

Kreiman, J., Gerratt, B. R., and Antonanzas-Barroso, N. *Analysis and Synthesis of Pathological Voice Quality (Software Manual)*. Available at (http://www.surgery.medsch.ucla.edu/glottalaffairs/software.htm), 2006.

Ladefoged, P. The nature of vowel quality. *Revista do Laboratorio de Fonetica Experimental da Faculdade de Letras da Universidade de Coimbra*, 5:73–162, 1960.

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

Ladefoged, P. and Maddieson, I. *The Sounds of the World's Languages*. Blackwell, 1996.

Laver, J. The description of voice quality in general phonetic theory. *Edinburgh university department of linguistics work in progress*, 12:30–52, 1979.

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

Laver, J. *Principles of Phonetics*. Cambridge University Press, 1994.

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

Lieberman, P. and Blumstein, S. E. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge University Press, 1988.

Lim, Il-Taek and Lee, Byeong Gi. Lossless pole-zero modeling of speech signals. *IEEE Transactions on Speech and Audio Processing*, 1(3):269–276, 1993.

Lin, Q. Nonlinear interaction in voice production. *KTH, Speech Transmission Laboratory, Quarterly Report*, 28(1):1–12, 1987.

Lin, Q. Speech production theory and articulatory speech synthesis. *Ph. D. Thesis, Royal Institute of Technology, Stockholm*, 1990.

Lliev, A. I., Scordilis, M., Papa, J., and Falcão, A. Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech and Language*, 24(3):445–460, 2010.

Luchsinger, R. and Arnold, G. *Voice - speech - language . Clinical communicology: its physiology and pathology*. Constable, London, 1965.

Lugger, M. and Yang, B. The relevance of voice quality features in speaker independent emotion recognition. *Proceedings of ICASSP, Honolulu, Hawaii*, 4:17–20, 2007.

Lugger, M., Yang, B., and Wokurek, W. Robust estimation of voice quality parameters under real world disturbances. *Proceedings of ICASSP, Toulouse, France*, pages 1197–1110, 2006.

Mackenzie Beck, J. Perceptual analysis of voice quality: The place of vocal profile analysis. In Laver, J., Hardcastle, W., and Mackenzie Beck, J., editors, *A figure of speech: A Festschrift for John Laver*, chapter 12, pages 285–322. 2005.

Magnuson, T. Realizations of /r/ in Japanese talk-in-interaction. *Proceedings of ICPhS, Hong Kong*, pages 1306–1309, 2011.

Mahshie, J. and Gobl, C. Effects of varying LF model parameters on KLSYN88 synthesis. *Proceedings of ICPhS, San Francisco*, pages 1009–1012, 1999.

Mallet, S. *A Wavelet Tour of Signal Procesing*. New York: Academic, 2nd edition, 1999.

Mallet, S. and Zhong, S. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7):710–732, 1992.

Markel, J. and Gray, A. *Linear Prediction of Speech*. Springer-Verlag, 1982.

Marwick, H., Mackenzie, J., Laver, J., and Trevarthen, C. Voice quality as an expressive system in mother-to-infant communication: A case study. *Edinburgh university department of linguistics work in progress*, 17:85–97, 1984.

Moisik, S. and Esling, J.H. The 'whole' larynx approach to laryngeal features. *Proceedings of ICPhS, Hong Kong*, pages 1406–1409, 2011.

Monoson, P. and Zemlin, W. Quantitative study of a whisper. *Folia Phoniatrica*, 36(2): 161–174, 1989.

Monsen, R. and Engebretson, A. Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, 62:981–93, 1977.

Moore, E., Clements, M., Peifer, J., and Weisser, L. Investigating the role of glottal features in classifying clinical depression. *Proceedings of the IEEE Conference: Engineering in Medicine and Biology Society*, 3:2849–2852, 2003.

Moore, E., Clements, M., Peifer, J., and Weisser, L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Transactions on Biomedical Engineering*, 55:96–107, 2008.

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

Murty, K. and Yegnanarayana, B. Epoch extraction from speech signals. *IEEE Transactions on audio speech and language processing*, 16:1602–1613, 2008.

Naylor, P., Kounoudes, A., Gudnason, J., and Brookes, M. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Audio Speech and Language processing*, 15(1):34–43, 2007.

Nelder, J. and Mead, R. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

Ney, H. Dynamic programming algorithm for optimal estimation of speech parameter contours. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:208–214, 1983.

Ní Chasaide, A. and Gobl, C. Contextual variation of the vowel voice source as a function of adjacent consonants. *Language and Speech*, pages 303–330, 1993.

Ní Chasaide, A., Yanushevskaya, I., and Gobl, C. Voice source dynamics in intonation. *Proceedings of ICPhS, Hong Kong*, pages 1470–1473, 2011.

Nord, L., Ananthapadmanabha, T. V., and Fant, G. Signal analysis and perceptual tests of vowel responses with an interactive source filter model. *KTH, Speech Transmission Laboratory, Quarterly Report*, (2-3):25–52, 1984.

Nord, L., Ananthapadmanabha, T. V., and Fant, G. Perceptual tests using an interactive source filter model and considerations for synthesis strategies. *Journal of Phonetics*, (14):435–442, 1986.

O' Cinnéide, A. *Phase distortion robust voice source analysis*. PhD thesis, Dublin Institute of Technology, 2012.

O' Cinnéide, A., Dorran, D., Gainza, M., and Coyle, E. A frequency domain approach to ARX-LF voiced speech parameterization and synthesis. *Proceedings of Interspeech, Florence*, pages 57–60, 2011.

Obin, N. Cries and whispers: Classification of vocal effort in expressive speech. *Proceedings of Interspeech, Portland, Oregon, USA*, 2012.

Ogarkova, A., Borgeaud, P., and Scherer, K. R. Language and culture in emotion research: a multidisciplinary perspective. *Social Science Information*, 48:339–357, 2009.

Ogden, R. Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association*, 31(1):139–152, 2001.

Ogden, R. The larynx, voicing and voice quality. In *An introduction to English phonetics*, pages 40–55. 2009.

Oppenheim, A. and Schafer, R. *Discrete-Time Signal Processing*. Prentice-Hall, 1989.

Ozdas, A., Shiavi, R., Silverman, S., Silverman, M, and Wilkes, D. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, 51(9):1530–1540, 2004.

Pantazis, Y. and Stylianou, Y. Improving the modeling of the noise part in the harmonic plus noise model of speech. *Proceedings of ICASSP, Las Vegas, Nevada, USA*, pages 4609–4612, 2008.

Plumpe, M., Quatieri, T., and Reynolds, D. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7(5):569–586, 1999.

Podesva, R. J. Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics,*, 11(4):478–504, 2007.

Pressman, J. J. Physiology of the vocal cords in phonation and respiration. *Archives of Otolaryngology*, 35:355–398, 1942.

Rabiner, R. and Schafer, R. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. HMM-based Finnish text-to-speech system utilizing glottal inverse filtering. *Proceedings of Interspeech, Brisbane, Australia*, pages 1881–1884, 2008.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Audio Speech and Language processing*, 19(1):153–165, 2011.

Rao, K.S. and Yegnanarayana, B. Prosody modification using instants of significant excitation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):972–980, 2006.

Rao, K.S., Prasanna, S.R.M., and Yegnanarayana, B. Determination of instants of significant excitation in speech using hilbert envelope and group delay function. *IEEE Signal Processing Letters,*, 14(10):762–765, 2007.

Rose, P. J. Phonetics and phonology of Yang tone phonation types in Zhenai. *Cahiers de Linguistique Asie Orientale*, 18:229–245, 1989.

Rosenberg, A. E. Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 49(2):583–590, 1971.

Rothenberg, M. Acoustic interaction between the glottal source and the vocal tract. In Stevens, K. and Hirano, M., editors, *Vocal fold physiology*, pages 305–323. University of Tokyo Press, Tokyo, 1981.

Sadanobu, T. Expressive speech and grammar. *Proceedings of the JST/CREST Workshop on Expressive Speech Processing*, pages 55–60, 2003.

Scherer, S., Schwenker, F., Campbell, N., and Palm, G. Multimodal laughter detection in natural discourses. *Human centered robot systems, Cognitive Systems Monographs*, 6: 111–120, 2009.

Scherer, S., Kane, J., Gobl, C., and Schwenker, F. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech and Language*, 2012a.

Scherer, S., Layher, G., Kane, J., Neumann, H., and Campbell, N. An audiovisual political speech analysis incorporating eye-tracking and perception data. *Proceedings of LREC, Istanbul, Turkey*, pages 1114–1120, 2012b.

Schnell, K. Estimation of glottal closure instances from speech signals by weighted nonlinear prediction. *Advances in Nonlinear Speech Processing*, pages 221–229, 2007.

Schroeder, M. Emotional speech synthesis: A review. *Proceedings of Eurospeech*, pages 561–564, 2001.

Schroeder, M. and Grice, M. Expressing vocal effort in concatenative synthesis. *Proceedings of ICPhS, Barcelona, Spain*, pages 2589–2592, 2003.

Schwarz, R. and Chow, Y-L. The N-best algorithm: an efficient procedure and exact procedure for finding the n most likely sentence hypotheses. *Proceedings of ICASSP, Albuquerque, New Mexico, USA*, pages 81–84, 1990.

Shue, Y. L. and Alwan, A. A new glottal source model based on high-speed imaging and its application to voice source estimation. *Proceedings of ICASSP, Dallas, Texas, USA*, pages 5134–5137, 2010.

Silen, H., Helander, E., Nurminen, J., and Gabbouj, M. Parameterization of vocal fry in HMM based speech synthesis. *Proceedings of Interspeech 2009, Brighton*, pages 1775–1778, 2009.

Silverman, D., Blankenship, B., Kirk, P., and Ladefoged, P. Phonetic structures in Jalapa Mazatec. *Anthropological Linguistics*, 37(1):70–88, 1995.

Slyh, R. E., Hansen, E. G., and Anderson, T. R. Glottal modeling and closed-phase analysis for speaker recognition. *Proceedings of the Speaker and Language recognition workshop, Toledo, Spain*, pages 315–322, 2004.

Smits, R. and Yegnanarayana, B. Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing*, 3(5):325–333, 1995.

Solomon, N. P., McCall, G. N., Trosset, M. W., and Gray, W. C. Laryngeal configuration and constriction during two types of whispering. *Journal of Speech and Hearing Research*, 24(5):574–584, 1989.

Stevens, K. and Hanson, H. Classification of glottal vibration from acoustic measurements. *Vocal Fold Physiology*, pages 147–170, 1994.

Strik, H. Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *Journal of the Acoustical Society of America*, 103(5): 2659–2669, 1998.

Strik, H., Cranen, B., and Boves, L. Fitting a LF-model to inverse filter signals. *Proceedings of Eurospeech 1993, Berlin*, pages 103–106, 1993.

Strube, H. W. Determination of the instant of glottal closure from the speech wave. *Journal of the Acoustical Society of America*, 56(5):1625–1629, 1974.

Sturmel, N., d'Alessandro, C., and Doval, B. A comparative evaluation of the zeros of z transform representation for voice source estimation. *Proceedings of Interspeech, Antwerp, Belgium*, pages 558–561, 2007.

Sturmel, N., d'Alessandro, C., and Rigaud, F. Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform. *Proceedings of ICASSP, Taipei, Taiwan*, pages 4517–4520, 2009.

Stylianou, Y. Synchronization of speech frames based on phase data with application to concatenative speech synthesis. *Sixth European Conference on Speech Communication and Technology*, pages 2243–2246, 1999.

Stylianou, Y. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 19(1):21–29, 2001.

Sun, R. and Moore, E. A preliminary study on cross-databases emotion recognition using the glottal features in speech. *Proceedings of Interspeech, Portland, Oregon, USA*, 2012.

Sun, R., Moore, E., and Torres, J. F. Investigating glottal parameters for differentiating emotional categories with similar prosodics. *Proceedings of ICASSP, Taipei, Taiwan*, pages 4509–4512, 2009.

Surana, K. and Slifka, J. Acoustic cues for the classification of regular and irregular phonation. *Proceedings of Interspeech (ICSLP), Pittsburgh, Pennsylvania, USA*, pages 693–699, 2006a.

Surana, K. and Slifka, J. Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English. *Proceedings of Speech Prosody, Dresden, Germany*, 2006b.

Székely, É., Cabral, J.P., Cahill, P., and Carson-Berndsen, J. Clustering expressive speech styles in audiobooks using glottal source parameters. *Proceedings of Interspeech, Florence, Italy*, pages 2409–2412, 2011.

Székely, É., Cabral, J.P., Abou-Zleikha, M., Cahill, P., and Carson-Berndsen, J. Evaluating expressive speech synthesis from audiobooks in conversational phrases. *Proceedings of LREC, Istanbul, Turkey*, pages 3335–3339, 2012a.

Székely, É., Kane, J., Scherer, S., Gobl, C., and Carson-Berndsen, J. Detecting a targeted voice style in an audiobook using voice quality features. *Proceedings of ICASSP, Kyoto, Japan*, pages 4593–4596, 2012b.

Tahon, M., Degottex, G., and Devillers, L. Usual voice quality features and glottal features for emotional valence detection. *Proceedings of Speech Prosody, Shanghai, China*, 2012.

Talkin, D. Voicing epoch determination with dynamic programming. *Journal of the Acoustical Society of America*, 85(S1):149–149, 1989.

Talkin, David. *A Robust Algorithm for Pitch Tracking*. Elsevier, 1995.

Teager, H. M. and Teager, S. M. Active fluid dynamic voice production models, or is there a unicorn in the garden. In Titze, I. and Scherer, R., editors, *Vocal fold physiology*, pages 387–401. 1983.

Teager, H. M. and Teager, S. M. Evidence for nonlinear sound production mechanisms in the vocal tract. In Hardcastle, W. J. and Marchal, A., editors, *Speech production and speech modelling*, pages 241–261. Kluwer Academic, 1990.

Thomas, M. and Naylor, P. The SIGMA algorithm: A glottal activity detector for electroglottographic signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1557–1566, 2009.

Thomas, M., Gudnason, J., and Naylor, P. Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing,*, 20(1):82–91, 2012.

Timcke, R., von Leden, H., and Moore, P. Laryngeal vibrations: measurements of the glottic wave. part 1: The normal vibratory cycle. *Archives of Otolaryngology - Head and Neck surgery*, 68(1):1–19, 1958.

Titze, I. Parameterization of the glottal area, glottal flow, and vocal fold contact areas. *Journal of the Acoustical Society of America*, 75:570–580, 1984.

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

Titze, I. and Sundberg, J. Vocal intensity in speakers and singers. *Journal of the Acoustical Society of America*, 91(5):2936–2946, 1992.

Traill, A. The laryngeal sphincter as a phonatory mechanism in !xóõ. In Singer, R. and Lundy, J. K., editors, *Variation, Culture and Evolution in African populations: Papers in Honour of Dr. Hertha de Villiers*, pages 123–131. Johannesburg: Witwatersrand University Press, 1986.

Tuan, V.N. and d'Alessandro, C. Robust glottal closure detection using the wavelet transform. *Sixth European Conference on Speech Communication and Technology*, pages 2805–2808, 199.

Vainio, M. *Artificial neural network based prosody models for Finnish text-to-speech synthesis*. PhD thesis, University of Helsinki, Finland, 2001.

Vainio, M., Airas, M., Järvikivi, J., and Alku, P. Laryngeal voice quality in the expression of focus. *Proceedings of Interspeech, Makuhari, Japan*, pages 921–924, 2010.

van den Berg, J. Myoelastic-aerodynamic theory of voice production. *Journal of speech and hearing research*, 1:227–244, 1958.

van den Berg, J. Mechanism of the larynx and the laryngeal vibrations. In Malmberg, B., editor, *Manual of phonetics*, pages 278–307. Amsterdam, 1968.

Van Riper, C. and Irwin, J. *Voice and Articulation*. Prentice-Hall, Englewood Cliffs, 1958.

Varga, A. and Steeneken, H.J.M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise

on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993. URL `http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html`.

Veeneman, D. E. and BeMent, S. L. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Trans. Acoust., Speech, Signal Proc*, 33:369–377, 1985.

Veldhuis, R. A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation. *Journal of the Acoustical Society of America*, 103(1):566–571, 1998.

Villavicencio, F., Robel, A., and Rodet, X. Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation. *Proceedings of ICASSP, Toulouse, France*, pages 869–872, 2006.

Vincent, D. *Analyse et controle du signal glottique en synthese de la parole (French)*. PhD thesis, ENST, Paris, France,, 2007.

Vincent, D., Rosec, O., and Chonavel, T. Estimation of lf glottal source parameters based on an arx model. *Proceedings of Interspeech, Lisbon, Portugal*, pages 333–336, 2005.

Vincent, D., Rosec, O., and Chonavel, T. A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and hnm modeling. *Proceedings of ICASSP*, pages 525–528, 2007.

Vishnubhotla, S. and Espy-Wilson, C. Automatic detection of irregular phonation in continuous speech. *Proceedings of Interspeech, Pittsburgh, USA*, pages 949–952, 2006.

Walker, J. and Murphy, P. A review of glottal waveform analysis. In Stylianou, Y., Faundez-Zanuy, M., and Esposito, A., editors, *Progress in nonlinear speech processing*, pages 1–21. Springer Verlag, 2007.

Warren, R. M. Acoustic repetition: Pitch and infrapitch. In *Auditory perception - A new synthesis*, pages 80–85. New York: Pergamon, 1982.

Wayland, R. and Jongman, A. Acoustic correlates of breathy and clear vowels: the case of Khmer. *Journal of Phonetics*, 31:181–201, 2003.

Wendahl, R. W. The roles of amplitude breaks in the perception of vocal roughness. *American speech and hearing association convention abstracts*, (6):406, 1964.

Wolk, L and Abdelli-Beruh, N. Habitual use of vocal fry in young adult female speakers. *Journal of Voice*, 26(3):111–116, 2012.

Wong, D., Markel, J., and Gray, A. Least squares glottal inverse filtering from acoustic speech waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27: 350–355, 1979.

Xiaochuan, Niu, Kain, A., and van Santen, J. P. Estimation of the acoustic properties of the nasal tract during the production of nasalized vowels. *Proceedings of Interspeech, Lisbon, Portugal*, pages 1045–1048, 2005.

Yanushevskaya, I., Gobl, C., and Ní Chasaide, A. Voice quality and f_0 cues for affect expression. *Proceedings of Interspeech, Lisbon, Portugal*, pages 1849–1852, 2005.

Yanushevskaya, I., Gobl, C., and Ní Chasaide, A. Voice parameter dynamics in portrayed emotions. *Proceedings of the 6th International Workshop: Models and analysis of vocal emissions for biomedical applications, Florence*, pages 21–24, 2009.

Yanushevskaya, I., Gobl, C., Kane, J., and Ní Chasaide, A. An exploration of voice source correlates of focus. *Proceedings of Interspeech, Makuhari, Japan*, pages 462–465, 2010.

Yanushevskaya, I., Ní Chasaide, A., and Gobl, C. Universal and language-specific perception of affect. *Proceedings of ICPhS, Hong Kong*, pages 2208–2211, 2011.

Yegnanarayana, B. and Veldhuis, R. Extraction of vocal-tract system characteristics from speech signals. *IEEE Transactions on Audio Speech and Language processing*, 6(4): 313–327, 1998.

Yu, K. M. and Lam, H. W. The role of creaky voice in Cantonese tonal perception. *Proceedings of ICPhS, Hong Kong*, pages 2240–2243, 2011.

Yuasa, I. K. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women? *American Speech*, 85(3):315–337, 2010.

Zelinka, P., Sigmund, M., and Schimmel, J. Impact of vocal effort variability on automatic speech recognition. *Speech Communication*, 54(6):732–742, 2012.

Zemlin, W. *Speech and Hearing Science*. Stipes Champaign, Illinois, 1964.

Zhang, C. and Hansen, J. Analysis and classification of speech mode: whisper through shouted. *Proceedings of Interspeech, Antwerp, Belgium,*, pages 2289–2292, 2007.

Zivanovic, M., Roebel, A., and Rodet, X. Adaptive threshold determination for spectral peak classification. *Proceedings of Conference on Digital Audio Effects (DAFx), Bordeaux, France*, pages 1–8, 2007.

Appendix A

GCI appendix

The pairwise comparisons following the statistical analysis described in Chapter 4 Section 4.5.6 are summarised in Tables A.1-A.2 (IR) and A.3-A.4 (IDA).

Table A.1: Summary of the pairwise comparisons for **Identification Rate (IR)** following ANOVAs and subsequent posthoc testing using Tukey's Honestly Significant Difference (HSD) test for modal, tense, breathy and harsh phonation types. *** ($p < 0.001$), ** ($p < 0.01$) and * ($p < 0.05$) indicate significant differences in favour of the algorithms on the vertical, while +++ ($p < 0.001$), ++ ($p < 0.01$) and + ($p < 0.05$) indicate significant differences in favour of the algorithms on the horizontal axis.

	ESPS	SEDREAMS	DYPSA	YAGA	ZFF
Modal					
SE-VQ	0.98	0.99	***	0.99	0.09
ESPS		0.94	***	1.00	0.41
SEDREAMS			***	0.95	0.05
DYPSA				+++	+++
YAGA					0.35
Tense					
SE-VQ	0.99	0.99	***	0.09	***
ESPS		0.99	***	0.43	*
SEDREAMS			***	0.15	***
DYPSA				+++	0.07
YAGA					0.21
Breathy					
SE-VQ	0.99	1.00	0.22	0.99	***
ESPS		0.99	0.58	0.99	***
SEDREAMS			0.22	0.99	***
DYPSA			0.22	0.28	***
YAGA					***
Harsh					
SE-VQ	0.99	0.98	*	0.99	***
ESPS		0.99	0.11	1.00	***
SEDREAMS			0.2	0.99	***
DYPSA				0.10	+
YAGA					***

Table A.2: Summary of the pairwise comparisons for **Identification Rate (IR)** following ANOVAs and subsequent posthoc testing using Tukey's Honestly Significant Difference (HSD) test for modal, tense, falsetto and creaky phonation types. *** ($p < 0.001$), ** ($p < 0.01$) and * ($p < 0.05$) indicate significant differences in favour of the algorithms on the vertical, while +++ ($p < 0.001$), ++ ($p < 0.01$) and + ($p < 0.05$) indicate significant differences in favour of the algorithms on the horizontal axis.

	ESPS	SEDREAMS	DYPSA	YAGA	ZFF
Falsetto					
SE-VQ	0.79	0.98	*	**	0.99
ESPS		0.36	***	***	0.92
SEDREAMS			***	*	0.92
DYPSA				+++	+++
YAGA					+++
Creaky					
SE-VQ	0.06	***	**	***	***
ESPS		0.48	0.76	***	0.36
SEDREAMS			0.99	0.06	0.99
DYPSA				*	0.99
YAGA					0.10

Table A.3: Summary of the pairwise comparisons for **Identification Accuracy (IDA)** following ANOVAs and subsequent posthoc testing using Tukey's Honestly Significant Difference (HSD) test for modal, tense, breathy and harsh phonation types. *** ($p < 0.001$), ** ($p < 0.01$) and * ($p < 0.05$) indicate significant differences in favour of the algorithms on the vertical, while +++ ($p < 0.001$), ++ ($p < 0.01$) and + ($p < 0.05$) indicate significant differences in favour of the algorithms on the horizontal axis.

	ESPS	SEDREAMS	DYPSA	YAGA	ZFF
Modal					
SE-VQ	0.94	0.78	***	0.24	0.99
ESPS		0.99	***	0.82	0.88
SEDREAMS			***	0.95	0.67
DYPSA				+++	+++
YAGA					0.16
Tense					
SE-VQ	0.99	*	***	***	0.99
ESPS		**	***	***	0.95
SEDREAMS			***	***	0.06
DYPSA				0.77	+++
YAGA					+++
Breathy					
SE-VQ	0.99	0.99	***	0.98	+++
ESPS		0.99	***	0.99	+++
SEDREAMS			***	0.99	+++
DYPSA				+++	+++
YAGA					+++
Harsh					
SE-VQ	0.98	0.95	**	0.21	0.99
ESPS		0.99	*	0.62	0.9
SEDREAMS			*	0.73	0.80
DYPSA				0.49	+++
YAGA					0.09

Table A.4: Summary of the pairwise comparisons for **Identification Accuracy (IDA)** following ANOVAs and subsequent posthoc testing using Tukey's Honestly Significant Difference (HSD) test for modal, tense, falsetto and creaky phonation types. *** ($p < 0.001$), ** ($p < 0.01$) and * ($p < 0.05$) indicate significant differences in favour of the algorithms on the vertical, while +++ ($p < 0.001$), ++ ($p < 0.01$) and + ($p < 0.05$) indicate significant differences in favour of the algorithms on the horizontal axis.

	ESPS	SEDREAMS	DYPSA	YAGA	ZFF
Falsetto					
SE-VQ	0.31	0.78	***	*	0.89
ESPS		*	***	***	0.92
SEDREAMS			**	0.48	0.17
DYPSA				0.54	+++
YAGA					+++
Creaky					
SE-VQ	*	0.99	***	0.91	0.99
ESPS		0.15	0.73	++	0.19
SEDREAMS			**	0.59	1.00
DYPSA				+++	++
YAGA					0.59

Appendix B

Voice analysis toolkit

As part of the contributions of this Ph. D. thesis I have made the algorithms developed publicly available here: https://github.com/jckane/Voice_Analysis_Toolkit/

Below is the README file contained in the toolkit, providing details of ownership, usage and relevant references.

```
########################################################################
########### VOICE ANALYSIS TOOLKIT #####################################
########################################################################

This software has been coded by John Kane at the Phonetics and Speech
Laboratory in Trinity College Dublin, early-2013. This work is supported
by the Science Foundation Ireland, Grant 07/CE/I1142 (Centre for Next
Generation Localisation, www.cngl.ie) and Grant 09/IN.1/I2631 (FASTNET).

The toolkit contains a range of matlab files for glottal source and
voice quality analysis. For most of the algorithms Matlab versions
since 2009 be suitable. However, for the CreakyDetection_CompleteDetection.m
function only Matlab versions since Matlab R2011a (and including the neural
network toolbox) will be able to run it. Note that the creak detection
algorithm used here was developed jointly by John Kane (Phonetics and
Speech Laboratory in Trinity College Dublin) and Thomas Drugman (University
of Mons, Belgium) and that same
algorithm is also available within the GLOAT toolkit.
```

```
GETTING STARTED

If requiring the use of the LF mode function the lf_Area_newton.c in the
general_fcns/ directory must be mex-ed, e.g., use this command:

mex lf_Area_newton.c

in the general_fcns/ directory.

Note however there mex-ed versions are already included for Linux (32-bit),
Mac (64-bit) and Windows (32-bit and 64-bit).

Note also that the SRH f0/VUV tracking algorithm from the GLOAT toolkit
(Thomas Drugman, University of Mons) is used with this toolkit. The GLOAT
toolkit can be downloaded from here:

GLOAT    - http://tcts.fpms.ac.be/~drugman/Toolbox/GLOAT.zip

To see details of the usage of each of the methods, simply use the
help command in matlab, e.g.,

help SE_VQ


TESTING

To do a test-run of the software on an included ARCTIC utterance,
use the command:

[x,fs]=wavread('arctic_a0007.wav');
test_Voice_Analysis_Toolkit(x,fs)


FEEDBACK

Note that the code here has been developed and tested in a UNIX based
```

```
environment. None of the directory handling has been hardcoded and hence
there should not be problems running this on a windows machine. Also,
the code has been developed for Matlab 2011b. There may be unforeseen
problems with previous (and potentially future) versions. If you have
any difficulties or issues with the code please email me: kanejo@tcd.ie


NOVEL ALGORITHMS - BRIEF DESCRIPTION

SE_VQ - Algorithm for detecting glottal closure instants. This is a further
        further development of the SEDREAMS algorithm (Drugman et al 2012)
        and is designed to improve selection of GCI candidates through the
        use of a dynamic programming algorithm. It also involves a
        post-processing step to remove false positives in creaky
        voice regions.

dyProg_LF  - Algorithm for fitting LF model pulses to an
        estimated glottal flow derivative waveform. The method involves
        an exhaustive search process using Rd, a dynamic programming algorithm
        to choose the optimal path of Rd values and a subsequent optimisation
        algorithm in order to refine the fit by varying all three R-parameters

MDQ  -  The maxima dispersion quotient (MDQ) is used for discriminating
        breathy to tense voice based on wavelet-based decomposition of the
        LP-residual signal. Dispersion is measured in the vicinity of the GCI

creak_detect - An algorithm for detecting creaky voice regions from speech
       signals. The detection is based on a combination of new and existing
       acoustic features relevant to creaky voice which are used as input
       features to a artificial neural networks based classifier, which
       has been trained on a wide range of speech. Note that the development of
       this method has been done in collaboration with Thomas Drugman,
       University of Mons, Belgium.
```

REFERENCES

Please refer to the relevant references below when using any of these algorithms in published studies.

Kane, J., Gobl, C., (2013) 'Evaluation of glottal closure instant detection in a range of voice qualities', Speech Communication 55(2), pp. 295-314.

Kane, J., Gobl, C. (2013) ''Automating manual user strategies for precise voice source analysis'', Speech Communication 55(3), pp. 397-414.

Kane, J., Yanushevskaya, I., Ni Chasaide, A., Gobl, C., (2012) Exploiting time and frequency domain measures for precise voice source parameterisation, Proceedings of Speech Prosody.

Kane, J., Gobl, C., (2013)''Wavelet maxima dispersion for breathy to tense voice discrimination'', IEEE Trans. Audio Speech & Language Processing, 21(6), pp. 1170-1179.

Kane, J., Gobl, C., (2011) 'Identifying regions of non-modal phonation using features of the wavelet tranform', Proceedings of Interspeech 2011.

Drugman, T., Kane, J., Gobl, C., 'Automatic Analysis of Creaky Excitation Patterns', Submitted to Computer Speech and Language.

Kane, J., Drugman, T., Gobl, C., (2013) 'Improved automatic detection of creak', 27(4), pp. 1028-1047, Computer Speech and Language.

Drugman, T., Kane, J., Gobl, C. (2012) Resonator-based creaky voice detection, Proceedings of Interspeech.

Drugman, T., Kane, J., Gobl, C. (2012) Modeling the creaky excitation for parametric speech synthesis, Proceedings of Interspeech.

##