Sādhanā Vol. 36, Part 5, October 2011, pp. 623–650. © Indian Academy of Sciences

Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications

PAAVO ALKU*

Department of Signal Processing and Acoustics, Aalto University, Otakaari 5 A,
P.O. Box 13000, 00076 Aalto, Finland
e-mail: paavo.alku@aalto.fi

Abstract. Glottal inverse filtering (GIF) refers to methods of estimating the source of voiced speech, the glottal volume velocity waveform. GIF is based on the idea of inversion, in which the effects of the vocal tract and lip radiation are cancelled from the output of the voice production mechanism, the speech signal. This article provides a review on GIF research by examining an era spanning five decades during which this topic has been under development. The topic is handled from three main perspectives: the estimation methods of the glottal source, the parameterization techniques that have been developed to express the estimated glottal excitations in numerical forms, and the application areas of GIF. Finally, the strengths and limitations of the GIF approach are discussed.

Keywords. Speech; voice; speech production; inverse filtering; glottis.

1. Introduction

Speech is one of the most fundamental phenomena in human cultures. It is such a self-evident part of our everyday life that we take it for granted and often forget that it is, in fact, the most developed means of communication in existence. Speech is also an extremely complex acoustical carrier of vocal information; it constitutes a combinatory system in which a small number of basic elements, phonemes or syllables, can be combined into an infinite number of different sequences such as words and phrases. Given the essential role that speech has in our society and its extreme complexity, it is natural that speech has become an important research goal in several scientific disciplines such as phonetics, linguistics, phoniatrics, cognitive neuroscience, and engineering. Research conducted in these areas of science has increased our understanding enormously about, for example, how humans produce and perceive speech and how different analysis methods can be used in order to extract knowledge from spoken language. Moreover,

*For correspondence

speech technology has had an essential role in the development of information and communications engineering that has affected extensively all levels of our society. In the past 20 years, this technological development has been particularly strong and advanced theoretical works have led to successful commercial products. These products are used in technologies such as speech *coding*, which is a methodology to compress digital voice signals for fast transmission and efficient storage, speech *enhancement*, speech and speaker *recognition* and speech *synthesis*.

All areas of speech science are, in one form or another, based on the utilization of the speech pressure waveform generated by the human voice production mechanism. Therefore, it can be argued that among the many areas of speech science, speech *production* has a special role and scientific understanding achieved in this area can be taken advantage of practically in all study areas of speech. Understanding the human voice production mechanism is, however, not only extremely difficult but also highly challenging due to the fact that humans are capable of varying extensively the functioning of their vocal organs. This dynamics in voice production implies that speech sounds perceived in our everyday spoken communication vary extensively depending on, for example, the utterance produced, the speaker, the language, and the emotional state of the speaker. A (grossly) simplified manner to study the functioning of the human speech production mechanism is to categorize speech sounds into three main classes according to the production mechanism (Flanagan 1972). The three classes are *voiced sounds*, which are excited by the fluctuation of the vocal folds, *unvoiced sounds*, where the sound excitation is turbulent noise, and *plosives*, which are transient-type sounds made up by abruptly releasing the air flow that has been blocked by, for example, the lips. Among these three categories, voiced sounds, and especially their sub-group, vowels, represent a study goal that has been of special interest in the history of speech science. The importance of voiced sounds, and especially that of vowels, is explained by their crucial phonetic role in most languages. Voiced sounds are also more frequent than unvoiced in most West European languages: in English, for example, 78% of speech sounds were reported by Catford (1977) to be voiced. From the acoustical point of view, the importance of vowels lies in their long duration and larger energy in comparison to many other utterances.

The excitation of voiced speech corresponds to the air flow that streams from the lungs and generates oscillations which take place in the mucosa of the vocal folds. This excitation is called the *glottal excitation*[1] after the orifice between the two vibrating vocal folds, the glottis (figure 1). The rate at which the vocal folds open and close is determined, to a large extent, by subglottal pressure. The glottal excitation is strongly filtered by the physiological filter, the vocal tract, that is made up by the vocal organs between the vocal folds and the lips. The filtering effects of the vocal tract can be influenced by the speaker, for example, by the positioning of the tongue, the degree of opening of the mouth and the movement of the lips. Consequently, the resonance frequencies of the vocal tract, which are referred to as the formants, vary and the glottal excitation is strongly affected in spectrum when filtered by the vocal tract. This time-varying physiological filter, characterized by its resonances, colours the glottal excitation with acoustical cues that carry phonemic information, which is utilized by the human speech perception mechanism to distinguish, for example, the vowel [a] from the vowel [i]. The final part in the (simplified) model of voiced speech production chain is the lip-radiation, which corresponds to changing the volume velocity signal at lips into a free-field speech pressure waveform (Fant 1970).

[1] The correct acoustical term for the excitation of voiced speech is *glottal volume velocity waveform*, which measures how the volume of air (in litres) passing through the glottis changes as a function of time (in seconds). In order to avoid such a long expression, this signal can also be named as *glottal flow, glottal excitation or glottal source*.

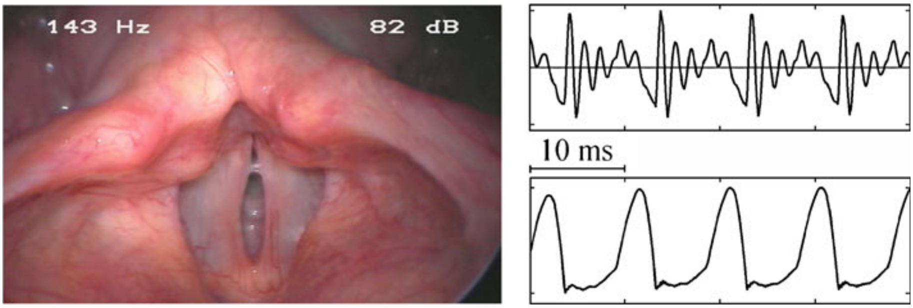

Figure 1. Left: A still image of vibrating vocal folds of a male speaker recorded by videolaryngoscopy at the Helsinki University Central Hospital (Helsinki, Finland). Right: Speech pressure waveform of a vowel sound (upper panel) and the corresponding glottal excitation (lower panel) estimated by an inverse filtering method described in Alku (1992).

Measurements of the vibrating vocal folds is difficult, as the oscillation can neither be examined directly due to the hidden position of the vocal folds nor be observed without aids because of the high speed of the oscillations. Instrumentation to carry out the measurements of the vibrating vocal folds has been developed during the recent decades and the methods developed are based, for instance, on visual (e.g., Sonesson 1959), electrical (e.g., Lecluse *et al* 1975) and electromagnetic methods (Titze *et al* 2000). Visual analysis is widely used especially in clinical investigation of voice production. Several techniques, such as video stroboscopy (e.g., Hirano 1981), digital high-speed stroboscopy (e.g., Eysholdt *et al* 1996) and kymography (Švec and Schutte 1996), have been developed, and many of them are currently used in daily practices in voice clinics. Acquiring visual information about voice production, however, always calls for invasive measurements in which the vocal folds are examined either with a solid endoscope inserted in the mouth or with a flexible fiberscope inserted in the nasal cavity. The use of these techniques might also be limited due to the fact that they need special equipment that is usually expensive.

As an alternative to the analysis methods referred to above, it is possible to investigate the functioning of the human speech production mechanism acoustically using *glottal inverse filtering* (GIF). The idea of GIF is to form a computational model for the filtering effects of the vocal tract and lip radiation (figure 2) and then to cancel these effects from speech by filtering the recorded signal through the *inverses* of the vocal tract and lip radiation model. In other words, GIF aims to estimate the input of the (vocal) system, the glottal excitation, when the output, the speech signal, is known. Hence, this analysis method introduces *an inverse problem* and from this perspective the study of voice production can be methodologically connected to a number of completely different areas of science such as geophysics, medical imaging and remote sensing. In most GIF methods, the estimation of the vocal tract and lip radiation can be computed solely from the acoustical speech pressure signal, which makes the analysis fully non-invasive. GIF is typically used in association with parameterization of the estimated glottal excitation in order to describe voice production quantitatively.

This article aims at providing a review on glottal inverse filtering by describing the topic from three perspectives. First, the GIF methods developed are described in chronological order. Second, the main parameterization methods that have been developed for the quantification of glottal excitation estimates are dealt with. Third, possible applications for GIF are discussed. The author is aware of three previous reviews on GIF (Hillman & Weinberg 1981; Fritzell 1992;

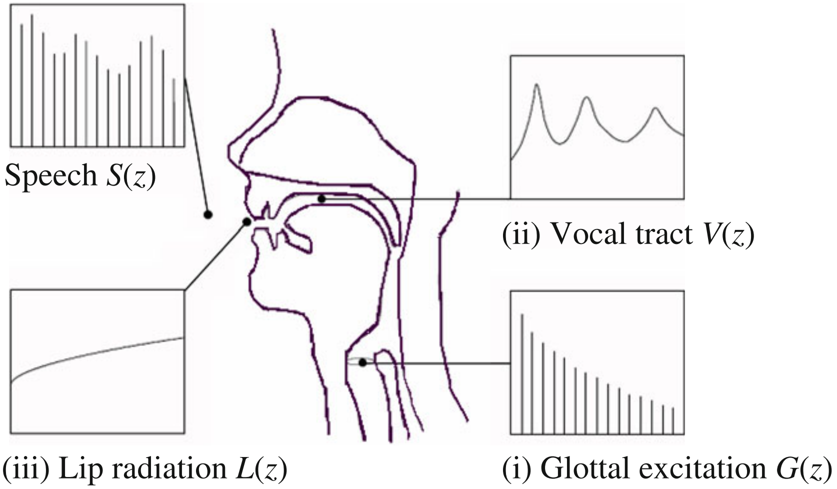

Figure 2. Schematic diagram describing the production of a speech pressure signal, denoted by S(z), as a cascade of three processes: (i) glottal excitation, (ii) vocal tract filtering, and (iii) lip radiation. Estimation of the glottal excitation with GIF corresponds to cancelling the filtering effects of the vocal tract and lip radiation from the speech signal:$G(z) = S(z) \cdot \frac{1}{L(z)} \cdot \frac{1}{V(z)}$.

Walker & Murphy 2005) and the current one, hopefully, adds new information to these previous reviews. Given the relatively long history (ca. 50 years) of the glottal inverse filtering research, it is simply not possible to discuss all published studies on this subject in this article. The review mainly handles studies which have been published in peer-reviewed journal articles, giving those presented in conference proceedings lesser weightage.

2. Glottal inverse filtering methods

2.1 *The 1950s and 1960s*

The article published by Miller in the late 1950s (Miller 1959) is widely considered to be the first study, which utilized the idea to compute an estimate for the glottal excitation by feeding the recorded speech signal through equipment that cancels the effect of the vocal tract. In these early days, all the equipment was naturally analogue and the inverse model of the vocal tract consisted of lumped elements. Only the first formant (F1) of the vocal tract was involved in Miller's pioneering study, and the frequency of the F1 was determined firstly by conducting a rough spectrographic analysis. After this, the fine-tuning of the resistors of the inverse network was conducted by searching for such values that yielded zero flow after the instant of glottal closure. The analysis also involved using an FM tape recorder from which the speech sound was fed into the inverse network, and an oscilloscope, which was used to visually analyse the obtained glottal excitation. Besides being the first GIF article, this work also described various phenomena of voice production, such as the occurrence of the main excitation of the vocal tract at the instant of glottal closure and the effect of increasing stress on glottal excitation, that have since been referred to in many articles.

Miller's work was followed by a series of GIF studies conducted in the early 1960s at KTH (Royal Institute of Technology, Stockholm, Sweden) especially by Gunnar Fant and Jan Lindqvist-Gauffin (Fant 1961; Fant & Sonesson 1962; Lindqvist-Gauffin 1963; Lindqvist-Gauffin 1964; Lindqvist-Gauffin 1965). The inverse filter used in these studies was also analogue but capable of handling four formants (F1–F4) instead of just F1 as in Miller's study. Again, the tuning of the anti-resonances was based on searching for such settings that yielded maximally flat closed phase of the glottal pulseform. This criterion was described by Lindqvist-Gauffin (1964) as follows: 'zeros are simply set to produce minimum formant ripples in the output waveform', which has later been used in several GIF studies as the major criterion in tuning the vocal tract inverse model. In parallel with studies conducted at KTH, similar inverse filter experiments were also conducted by Holmes (1963) and Carr and Trill (1964). Mathews *et al* (1961) proposed a pitch-synchronous analysis technique, which was used a decade later by Rosenberg (1971) in glottal inverse filtering. In this technique, the Fourier transform is computed over a fundamental period of the speech pressure signal and the formant frequencies are determined using a trial spectrum, which is fit to the observed spectrum.

Although GIF studies in the 1950s and 1960s were characterized by analogue techniques, the first experiments utilizing digital signal processing in the estimation of the glottal source were published towards the end of 1960s. Oppenheim and Schafer (1968) studied the so-called homomorphic analysis of speech in which the convolved components of voice are transformed into additive components using the cepstral analysis. These authors used the homomorphic analysis as a digital inverse filtering method in which the effects of the vocal tract are cancelled by retaining only that part of the cepstrum where the time index is less than the pitch period. Digital cepstral analysis has been later used by, e.g., Drugman *et al* (2009) as a means to separate the glottal source from the vocal tract.

2.2 *The 1970s and 1980s*

The use of digital filters to implement anti-resonances in the GIF analysis was investigated in the beginning of the 1970s by Nakatsui and Suzuki (1970). Their study, which has unfortunately remained largely unnoticed in later investigations, utilized five digital filters for anti-resonances for the vocal tract. The first three of these were adjustable and the remaining two were fixed. Although the study by Nakatsui and Suzuki is concise and their results on natural speech are limited, this study has its special role among GIF studies because it introduced digital filtering as a means to implement the anti-resonances of the vocal tract inverse model.

In 1973, Rothenberg introduced a new glottal inverse filtering technique that is one of the most well-known study (Rothenberg 1973). The proposed method is based on inverse filtering the volume velocity waveform recorded in the oral cavity rather than the radiated acoustic speech pressure wave captured in the free field outside the mouth. Rothenberg designed a special pneumotachograph mask, later widely referred to as the Rothenberg's mask, which is a transducer capable of measuring the volume velocity at the mouth with a reasonable accuracy. The recorded signal was then subjected to inverse filtering, where analogue antiresonances for the lowest two formants were determined using a spectrographic analysis. Although the bandwidth of the analysis method is limited to ca. 1.7 kHz (Hertegård & Gauffin 1992), the estimated glottal excitation can be calibrated to correspond to absolute flow values including the DC flow, a significant feature that has since been taken advantage of in various voice production studies of normal and pathological voices (see references in section 3). As sequels to his classical paper published in 1973, Rothenberg later studied, for example, how the pneumotachograph mask can be optimized

(Rothenberg 1977) and how the mask recordings can be used in the estimation of the glottal area function by utilizing time-varying non-linear inverse filtering (Rothenberg & Zahorian 1977). It is also worth noting that although the original work of Rothenberg involved the use of analogue filters in cancelling the vocal tract resonances, the use of the pneumotachograph mask was later combined with digital inverse filtering (e.g., Granqvist *et al* 2003).

In addition to Rothenberg's work, a significant progress was achieved in GIF methods in the mid and late 70s in the form of a technique that is known as the closed phase covariance method (Strube 1974; Wong *et al* 1979). This method can be regarded as the first such *digital* inverse filtering algorithm that is both widely known and applied in a wide range of studies (see references in section 3). The method uses linear prediction (LP) with the covariance criterion (Rabiner & Schafer 1978) as a tool to compute a digital all-pole model for the vocal tract. The use of LP introduces a remarkable improvement in the computation of the inverse model of the vocal tract because the model adjusts automatically to the underlying speech signal. This implies that time-consuming and clumsy adjustments of antiresonances used in analogue GIF methods can be avoided. The LP model of the vocal tract in the closed phase analysis is computed from the speech pressure signal over a time span when there is no contribution from the excitation, that is, from the samples that occur during the closed phase of the glottal cycle. The closed phase analysis undoubtedly provides a more practical analysis method of voice production than its older analogue rivals. In addition, several studies have shown that the closed phase analysis yields accurate estimates of the glottal excitation for normal speech with low fundamental frequency and well-defined closed phase (e.g., Veeneman & BeMent 1985; Krishnamurthy & Childers 1986). A series of studies has, however, demonstrated that the method is particularly sensitive to the extraction of the closed phase position, and even small errors might result in severe distortion of the estimated glottal excitation (e.g., Larar *et al* 1985; Riegelsberger & Krishnamurthy 1993; Yegnanarayana & Veldhuis 1998). This drawback can be partly alleviated by using a two-channel approach, where the instant of glottal closure is extracted from the electroglottography (EGG) instead of the acoustic speech signal (e.g., Krishnamurthy & Childers 1986). An example of a glottal excitation estimated with this approach is shown in figure 3. The estimation accuracy of the closed phase analysis remains, however, poor if the duration of the closed phase is short, which occurs in high-pitch speech or sounds produced using a breathy phonation type. Nevertheless, the introduction of the closed phase analysis and especially its computational vocal tract modelling tool, the LP analysis, in the mid and late 70s can be interpreted as a change of generation from old analogue, user-adjustable systems to modern digital systems, where antiresonances of the vocal tract model could be automatically extracted from the digital speech wave. However, there are also studies published at the same time (e.g., Hunt *et al* 1978) that aim at combining the advantages of both analogue and digital techniques in the form of interactive digital GIF. The interactive technique proposed by Hunt *et al* (1978) has later been used together with EGG (Hunt 1987).

Development of digital GIF techniques was continued in the 1980s by utilizing the LP analysis in one form or the another. An example of this era is the study by Matausek and Batalov (1980), who proposed a two-stage procedure for the estimation of the glottal excitation. In their method, a first version of the glottal source was computed by utilizing the LP analysis with the covariance criterion to the pre-emphasised speech signal. A pole-zero glottal model was then fitted to the obtained signal and the final estimate of the (time-reversed) glottal source was obtained by exciting this filter with an impulse train. Further examples of the GIF studies conducted in the 1980s are the studies by Veeneman & BeMent (1985) and Krishnamurthy & Childers (1986), which addressed the problematic extraction of the closed phase position by taking advantage of EGG. In both these studies, the importance of automatic analysis was emphasized. A different

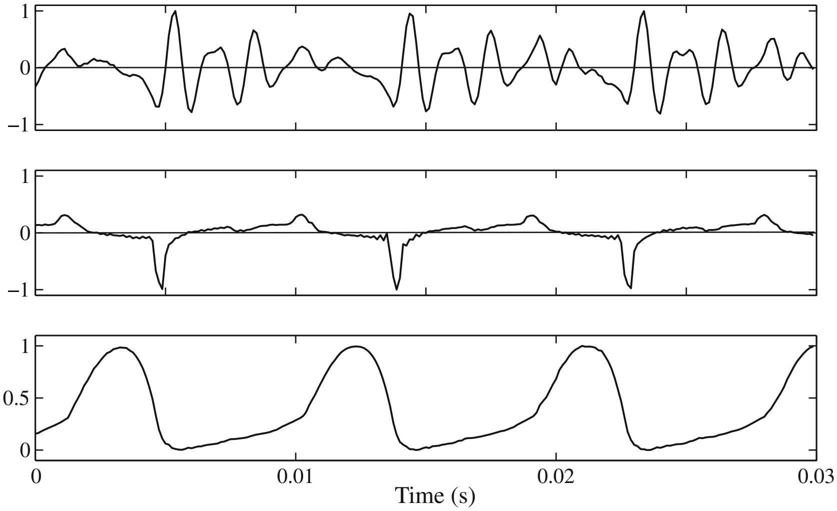

Figure 3. Estimation of the glottal excitation with the CP-analysis by extracting the position of the closed phase from the differentiated EGG. Waveforms shown: speech pressure signal to be analysed (top panel), differentiated EGG (middle panel), and the estimated glottal excitation (bottom panel). The closed phase was estimated to span between the instants of the minimum and maximum of the differentiated EGG. All signals are shown on arbitrary amplitude scales.

approach was studied by Milenkovic (1986), who proposed an idea, where a parametric all-pole model of the vocal tract is optimized *jointly* with a linear source model. The motivation behind this idea was to enable utilizing speech samples over the entire fundamental period in the optimization of the vocal tract, an issue which is not fulfilled in the closed phase covariance method. This approach was adopted by Milenkovic to help in the estimation of the glottal excitation from high-pitch speech and in sounds where the glottal closure is not complete. The idea of the joint optimization was later used during the same decade (e.g., Fujisaki & Ljungqvist 1987; Isaksson & Millnert 1989) and especially in the 1990s and 2000s by several other researchers as described in the following sub-section. Finally, the use of GIF as an analysis tool for linguistic research was addressed in the study by Javkin *et al* (1987). Their GIF method was digital, based on finding vocal tract resonances with LP, but instead of aiming at a fully automatic approach, the method involved a stage, where formant data given by LP was edited for errors, such as formant discontinuities, by a human experimenter.

The 1970s and 1980s were an era during which several major GIF studies were published as indicated by the studies referred to above. One more study from this era that needs to be mentioned in this context is by Sondhi (1975), who proposed a completely different idea to measure the glottal excitation. This technique, often named as the *Sondhi tube*, is based on reducing the effect of vocal tract resonances by speaking into a reflectionless tube rather than using a specific piece of equipment to implement antiresonances. By replacing the open end of the vocal tract at lips with a reflectionless tube, the method sets up such acoustic conditions that the vocal tract transmits the glottal pulse with minimal distortion. If the termination is perfect, the glottal excitation can be simply recorded by a probe microphone placed in the tube. In principle, this method is not a true GIF technique because the effects of the vocal tract are not cancelled by utilizing an antiresonance filter. The Sondhi tube, however, constitutes an alternative to GIF techniques and it has been used, as shown in later sections, in similar applications as, for instance, the CP technique.

2.3 *The 1990s and 2000s*

A straightforward and automatic GIF method was proposed by Alku in 1992. This method estimates the contribution of the glottal excitation on the speech spectrum with a low order LP model that is computed with a two-stage procedure. The vocal tract is then estimated using either conventional LP or discrete all-pole modelling (DAP). The latter is a parametric spectral modelling method, which, in comparison to the conventional LP, yields more accurate formant estimates in high-pitch sounds (El-Jaroudi & Makhoul 1991; Alku & Vilkman 1994). The GIF method proposed by Alku (1992) has been evaluated using physical modelling approaches (Alku *et al* 2006b; Alku *et al* 2006c) and is available as a free software package (Airas 2008). Kasuya and his co-authors studied the joint estimation of the glottal excitation and vocal tract (Ding *et al* 1997; Kasuya *et al* 1999). Their approach was based on the autoregressive model with an exogenous (ARX) input, where the input was represented by the Rosenberg–Klatt (Klatt & Klatt 1990) model of the glottal source. The technique also involved quantifying turbulent noise from voiced speech segments. Shapira & Gath (1998) proposed a completely different method to estimate the glottal excitation based on the fuzzy clustering of hyperplanes. Their technique is strictly speaking not an inverse filtering method, but it is meant to be used for the same purpose, the estimation of the glottal excitation from an acoustic speech signal, and is, therefore, an interesting alternative to traditional GIF techniques. At the end of the 1990s, a comparison of GIF methods utilizing the Rothenberg's mask was made by Södersten *et al* (1999). Their study provides a well-documented description of how inverse filtering parameters are adjusted when analysing challenging data represented by speech of varying loudness produced by females.

In the early 2000s, an automatic GIF method was developed by Fröhlich *et al* (2001). Their method uses DAP as a parametric model of the vocal tract and the Liljencrants–Fant (LF) model (Fant *et al* 1985) to simulate the derivate of the voice source. The best set of the LF parameters are searched for in an iterative procedure that utilizes multi-dimensional optimization techniques. Joint estimation of the glottal source and vocal tract based on the ARX model was conducted in the studies of Fu and Murphy (2003; 2006). Unlike Kasuya's studies (1999), these authors used the LF model instead of the Rosenberg–Klatt model as a parametric representation of the glottal source. A new idea for the estimation of the glottal excitation was proposed in the studies by Bozkurt and his co-authors (Bozkurt *et al* 2005; Bozkurt *et al* 2007; Sturmel *et al* 2007; Drugman *et al* 2009). Their method, zeros of Z-transform (ZZT), does not utilize the LP analysis in the estimation of the source-tract separation, but rather expresses the speech sound with the help of the z-transform as a large polynomial. The roots of the polynomial are separated into two patterns, corresponding to the glottal excitation and the vocal tract, based on their location from the unit circle. Akande and Murphy (2005) addressed several problems underlying the utilization of conventional LP in GIF such as estimation of the glottal contribution from the speech spectrum. As a solution, they proposed an algorithm, where the glottal contribution is first suppressed from the speech with frequency-selective multi-pole high-pass filtering. High-pass filtering is implemented using an adaptive all-pole filter whose coefficients and order are adjusted to minimize the effect of the peak of the glottal source spectrum, the so-called glottal formant, on the first formant. To achieve the zero-phase lag response, the input speech signal is first forward filtered and then backward filtered with this all-pole filter. Covariance analysis is then used to estimate the vocal tract transfer function, but the order of the model is selected adaptively based on the minimum phase criterion. The authors argue that the proposed GIF method shows clear improvement when compared to the conventional closed phase analysis.

The earlier sections have covered some of the main studies conducted on GIF techniques from the end of the 1950s onwards. In addition, there are several journal articles, conference proceeding papers and other reports on GIF methods that cannot be described in detail herein and are referred to in the form of a citation list. The author seeks to direct the reader's attention at least to the following methodological GIF studies: Holmes (1976), Boves & Cranen (1982), Cranen & Boves (1988), Thomson (1992), Childers *et al* (1995), Lu & Smith (1999), Arroabarren & Carlosena (2003), Shiga & King (2003), Moore & Clements (2004), Deng *et al* (2006), Schnell & Lacroix (2007), Dalsgaard *et al* (2008), Gudnason *et al* (2009) and Perez & Bonafonte (2009).

3. Parameterization methods of the glottal excitation

GIF analysis is typically followed by another stage, the parameterization of the estimated glottal excitations, where the obtained waveforms are expressed numerically with properly selected quantities. These quantities, the glottal flow parameters, aim at representing the most important features of the computed waveforms in a compressed numerical form. Therefore, the selection of the parameterization method is an essential part of voice production research because it reflects the transfer of information from the estimated excitation waveforms to the experimenter and applications. Successful GIF analysis of human voice production, therefore, requires knowledge of the different alternatives available in the parameterization stage and the kind of behaviour in the glottal function they focus on. With this knowledge, it is possible to select the best possible numerical measure to represent the estimated glottal excitations for the research problem in question.

A large number of different methods have been developed for the parameterization of the glottal flow. These techniques are discussed here by dividing the methodologies into two categories: time-domain and frequency-domain methods.

3.1 *Time-domain parameterization methods*

3.1a *Time-based measures*: One of the most widely used methods to parameterize glottal excitations is to define the simple time-based quantities described in figure 4. These are all defined by extracting (either manually or automatically) critical time-instants (glottal opening and closure, and the instant of maximal flow) from the excitation waveforms. These classical time-based measures were originally proposed by Timcke *et al* (1958), who defined the open quotient (OQ) and speed quotient (SQ). Later, Monsen & Engebretson (1977) proposed the closing quotient (ClQ). OQ is sometimes replaced with its complement, the closed quotient (e.g., Iwarsson *et al* 1998; Sundberg *et al* 1999a, Sundberg *et al* 1999b). In addition to these measures, it is also possible to quantify the glottal function by utilizing the time-domain features from the derivative of the source waveform, which was used, for example, by Price (1989), who measured the return quotient.

Computation of the classical time-based quantities from glottal excitations estimated from natural speech is often problematic due to formant ripple (i.e., the fluctuating component embedded in the estimated glottal excitation due to incomplete cancelling of formants by the inverse filter). In addition, the computed glottal source estimates might contain noise which originates either from the recording environment or from the voice production mechanism itself (e.g., aspiration noise). Even in the absence of formant ripple or noise, computation of OQ and SQ is difficult

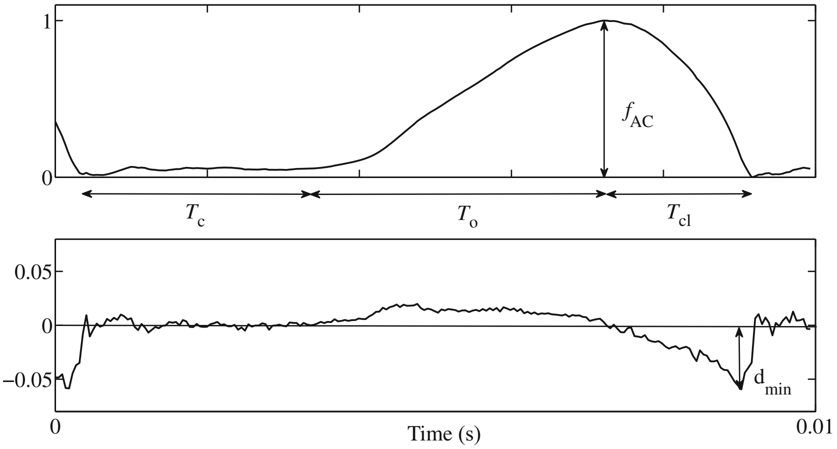

Figure 4. Computation of time-based parameters from the glottal pulse (upper panel) and its first time-derivative (lower panel). Time-durations marked: closed phase (T_c), opening phase (T_o), and closing phase (T_{cl}). Classical time-based parameters, open quotient (OQ), speed quotient (SQ), and closing quotient (ClQ) are defined from these values as follows: $OQ = \frac{T_o+T_{cl}}{T}$, $SQ = \frac{T_o}{T_{cl}}$, and $ClQ = \frac{T_{cl}}{T}$. Computation of the normalized amplitude quotient (NAQ) is defined from two amplitude-domain values, the AC-flow (f_{AC}), and the minimum of the derivative (d_{min}), as follows: $NAQ = \frac{f_{AC}}{d_{min} \cdot T}$. In all computations, the duration of the fundamental period is denoted by T ($T = T_c + T_o + T_{cl}$).

because of the gradual opening of the vocal folds. Therefore, computation of the time-based parameters is sometimes performed by replacing the true time instants of the glottal opening and closure by the time instants when the glottal flow crosses a level which is set to a certain ratio (e.g., 50%) of the difference between the maximum and minimum amplitude of the glottal cycle (Dromey *et al* 1992; Sapienza *et al* 1998).

It is also worth noting that the glottal closure instant, which is used above as one of the critical time-instants in defining the classical time-based parameters, can also be estimated from speech signals without using GIF. Estimation of glottal closure instants is widely referred to as the epoch extraction of speech (Ananthapadmanabha & Yegnanarayana 1975; Smits & Yegnanarayana 1995; Murty & Yegnanarayana 2008). The developed methods can be used, for instance, in the estimation of fundamental frequency and in modifying the prosodic structure of speech signals (Rao & Yegnanarayana 2006; Yegnanarayana & Murty 2009).

To simplify the extraction of the time-based parameterization, it is also possible to quantify the time-based features of the glottal source by measuring the amplitude-domain values as studied by Fant and his co-authors (Fant & Lin 1988; Fant *et al* 1994; Fant 1995; Fant 1997) as well as by Alku and his co-authors (Alku & Vilkman 1996b; Alku *et al* 2002; Bäckström *et al* 2002; Alku *et al* 2006a). One such measure is represented by the NAQ (normalized amplitude quotient) parameter proposed by Alku *et al* (2002). This parameter is computed from two amplitude-domain values, the AC-amplitude of the flow and the negative peak amplitude of the glottal flow derivative (see figure 4). It is worth noticing that these two amplitude measures are the extreme values of the flow and its derivative and, therefore, they are straightforward to be extracted from the output of GIF even though the estimated glottal source would be distorted by formant ripple. It can be shown that the ratio of these two amplitude-domain values is a time-domain quantity, which is interpreted by Fant as ‘the projection on the time axis of a tangent to the glottal flow at the point of excitation, limited by ordinate values of 0 and the AC-amplitude of the flow' (Fant 1997). It was shown by Alku *et al* (2002) and Bäckström *et al* (2002) that there is a high correlation between NAQ and ClQ. However, NAQ was shown to be more robust to noise present in the glottal flows obtained by inverse filtering natural speech and this improved robustness was largest in normal and pressed voices. Although, NAQ is a counterpart of ClQ defined from amplitude-domain measures, other classical time-based measures can also by represented similarly as shown by Gobl & Ní Chasaide (2003a).

3.1b *Amplitude-based measures*: When inverse filtering is done with a properly calibrated Rothenberg's mask it is possible to obtain valuable amplitude-based information from the glottal flow and its derivative. Figure 5 shows the amplitude-domain measures that are typically extracted from the glottal flow (upper panel) and its derivative (lower panel). The most widely used amplitude parameters are minimum flow (also called the DC-offset), the AC-flow and the negative peak amplitude of the flow derivative (also called the maximum airflow declination rate). In addition, the ratio between the AC and DC information has been used in the amplitude-based quantification of the glottal source (Isshiki 1981). Mean values for the minimum flow in speech produced using normal loudness have been reported to be ca. 0.10 l/s for both genders, although the values for males are slightly larger.

All the parameterization methods described so far are based on extracting certain time-based or amplitude-based measures from the most important instants of the flow or from its derivative. It is also possible to parameterize the voice source by searching for an artificial waveform that matches the computed glottal excitation (or its derivative). In other words, one aims to look for a compressed set of parameters that model the entire flow (or its derivative) instead of measuring isolated critical instants of the waveform. This approach implies selecting a set of pre-defined mathematical functions to model the glottal flow (or its derivative) and a procedure, which optimizes the parameters of the function in order to get the best possible match between the waveform given by inverse filtering and its artificial counterpart. There are many such artificial glottal source models that have been developed during the past three decades. One of the

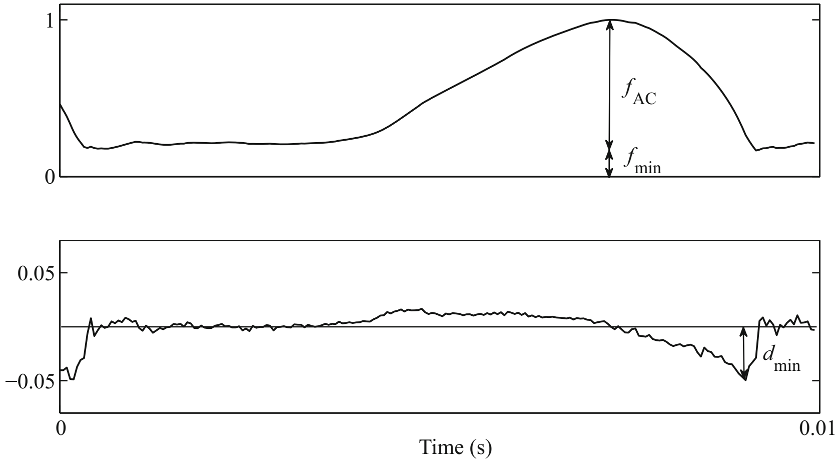

Figure 5. Computation of amplitude-domain parameters from the glottal pulse (upper panel) and its first time-derivative (lower panel). AC-flow (f_{AC}), minimum flow ($f_{\min}$), and the minimum of the derivative ($d_{\min}$).

most straightforward models is the two-parameter[2] polynomial model proposed by Klatt (Klatt 1987). In this model, the time-domain waveform of the glottal flow pulse is expressed during the glottal open phase with a third order polynomial given in Eq (1).

$$u_{\mathrm{g}}(t) = a \cdot t^2 - b \cdot t^3. \tag{1}$$

In Eq (1), parameters a and b are defined in order to obtain a glottal pulse with a desired OQ. Examples of more complicated glottal source models are the Rosenberg model (Rosenberg 1971), the Hedelin model (Hedelin 1984), and the Fujisaki–Ljungqvist model (Fujisaki & Ljungqvist 1986). The most widely used artificial model of the (differentiated) glottal excitation is the Liljencrants–Fant model (LF-model) (Fant *et al* 1985). In this model, the flow derivative is presented in two segments. The first segment models the flow derivative from the glottal opening (t_{o}) until the instant of the main excitation (t_{e}) with an exponential and sinusoidal function according to Eq (2).

$$E(t) = u'_{\mathrm{g}}(t) = E_0 \cdot e^{\alpha \cdot t} \sin(\varpi_{\mathrm{g}} t). \tag{2}$$

For the second segment, the flow derivative is modelled from the instant of the main excitation to the instant of (full) closure with exponential functions according to Eq (3).

$$E(t) = u'_{\mathrm{g}}(t) = \frac{-\mathrm{EE}}{\varepsilon \cdot T_{\mathrm{a}}} \left(e^{-\varepsilon \cdot (t - t_{\mathrm{e}})} - e^{-\varepsilon \cdot T_{\mathrm{b}}} \right). \tag{3}$$

In addition to critical time instants t_{o} and t_{e}, Eqs (2) and (3) have altogether seven parameters: E_0 (a scaling factor that is used to achieve the area balance between the opening and closing phases), α (rate of the amplitude increase), ω_{g} (sine frequency), EE (amplitude of the negative peak), ε (inverse of the time-constant of the exponential function of the return phase), T_a (effective duration of the return phase), and T_b (duration of the return phase). The LF model is, however, referred to as a four-parameter model, with F0 as the fifth parameter, because the number of parameters can be reduced by iteration and by requiring the integral of the pulse over the glottal cycle to be zero. For more details about the computation of the LF parameters, the reader is referred to the introduction in Gobl (2003). Computationally more effective versions of the LF model have been later developed by, for example, Veldhuis (Veldhuis 1998). It is also possible to fit the voice source over a single glottal cycle by using polynomials (Childers & Hu 1994; Kaburagi & Kawai 2003). In this case, however, the model parameters have no such physiological correspondence as, for example, in the LF parameters. Time-domain reproduction of the glottal excitation estimated by inverse filtering is also possible by taking advantage of physical models of voice production (Avanzini *et al* 2001; Drioli 2005; Avanzini 2008). This approach is in principle different from the use of mathematical functions in mimicking the voice source, which is done, for example, in the LF model, because the technique aims at searching for the physical parameters (e.g., vocal fold mass and stiffness) of the underlying oscillator rather than explaining the waveform of the volume velocity signal (or its derivative).

Finally, it is worth mentioning that amplitude-domain information can also be utilized using the phase-plane domain (Edwards & Angus 1996; Bäckström *et al* 2005, Moore & Torres 2008). In this approach, the glottal excitation is expressed in a two-dimensional space spanned by the

[2] Artificial models of the glottal pulse use a different number of parameters. The length of the pulse is typically not involved in the parameter set even though it is naturally needed in order to synthesise a glottal pulseform with a desired F0. For the Klatt waveform, this implies that the true number of parameters needed is three (i.e., two parameters to modify the shape of the pulse plus one additional parameter to define the duration of the waveform).

amplitude of the waveform and the amplitude of the derivative of the waveform. Although this method has not been developed strictly speaking for the parameterization purposes but rather for the assessment of the quality of inverse filtering, it can, in principle, be used as a means of visualizing the amplitude features of the glottal excitation.

3.2 *Frequency-domain parameterization methods*

Humans are able to vary the characteristics of the glottal pulse form by regulating the tension of the laryngeal muscles together with respiratory effort. The waveform of the volume velocity pulse can vary from a smooth, symmetric form, which is typical of soft sounds, to an asymmetric waveform with sharp edges that occurs, for example, in production of loud voices. This kind of time-domain variation is also reflected in the frequency domain as an alternation in the decay of the spectral envelope of the glottal pulse as shown in figure 6. Therefore, it is natural that parameterization of the glottal excitation has also been approached by utilizing frequency-domain measures, especially those focusing on the quantification of the spectral decay of the glottal source. The source spectrum is typically computed using the fast Fourier transform (FFT). In most of the cases, the spectrum is computed pitch-asynchronously over several fundamental periods and the spectral decay is measured from the levels of the harmonics. In addition, it is possible to quantify the spectral decay of the glottal source by using the FFT that is computed pitch-synchronously over a single glottal cycle. Finally, as an alternative to non-parametric spectral measures such as the FFT, it is possible to fit the glottal excitation with a parametric spectral model using, for example, auto-regressive (AR) spectral estimation methods.

The most straightforward parameterization method for the spectral skewness is the alpha ratio, which is the ratio between spectral energies below and above a certain frequency limit (typically 1.0 kHz) (Frøkjær-Jensen & Prytz 1973). (There is some discrepancy in the definition of the alpha ratio: the frequency limits might vary and sometimes the measure is defined as the ratio between the energies above and below the limit). Measuring the spectral decay of the glottal source is, however, more justified to be computed by utilizing the level of the fundamental frequency (F0) and its multiple integers, the harmonics. This kind of measure was developed by Childers & Lee (1991), who presented a quotient, called harmonic richness factor (HRF), which is defined from the spectrum of the estimated glottal flow as the ratio between the sum of the amplitudes of harmonics above the fundamental and the amplitude of the fundamental. Applying the spectral harmonics of the glottal airflow waveform in the quantification of voice production has also been used by Howell & Williams (1988; 1992), who measured the decay of the voice source spectrum by computing linear regression analysis over the first eight harmonics. Titze & Sundberg (1992) analysed the spectral decay of the voice source of singers by computing the difference between the amplitude of the fundamental and the second harmonic. This measure, usually denoted by H1–H2, is demonstrated in figure 6 by two examples. A measure of the voice source spectrum based on a pitch-synchronously computed spectrum was presented by Alku *et al* (1997). This measure, the parabolic spectral parameter (PSP), matches a second-order polynomial to the flow spectrum computed over a single glottal cycle and has been shown to be able to differentiate phonation types of varying spectral slopes effectively. In addition to measuring the spectral decay, it is also possible to utilize FFT-based techniques, such as the harmonic-to-noise ratio, to estimate the ratio between the harmonic and non-harmonic components (e.g., Murphy 1999). This approach is justified especially in the analysis if disordered voices, in which the glottal excitation typically involves an increasing amount of aperiodicities due to jitter, shimmer, aspiration noise and changing of the pulse waveform.

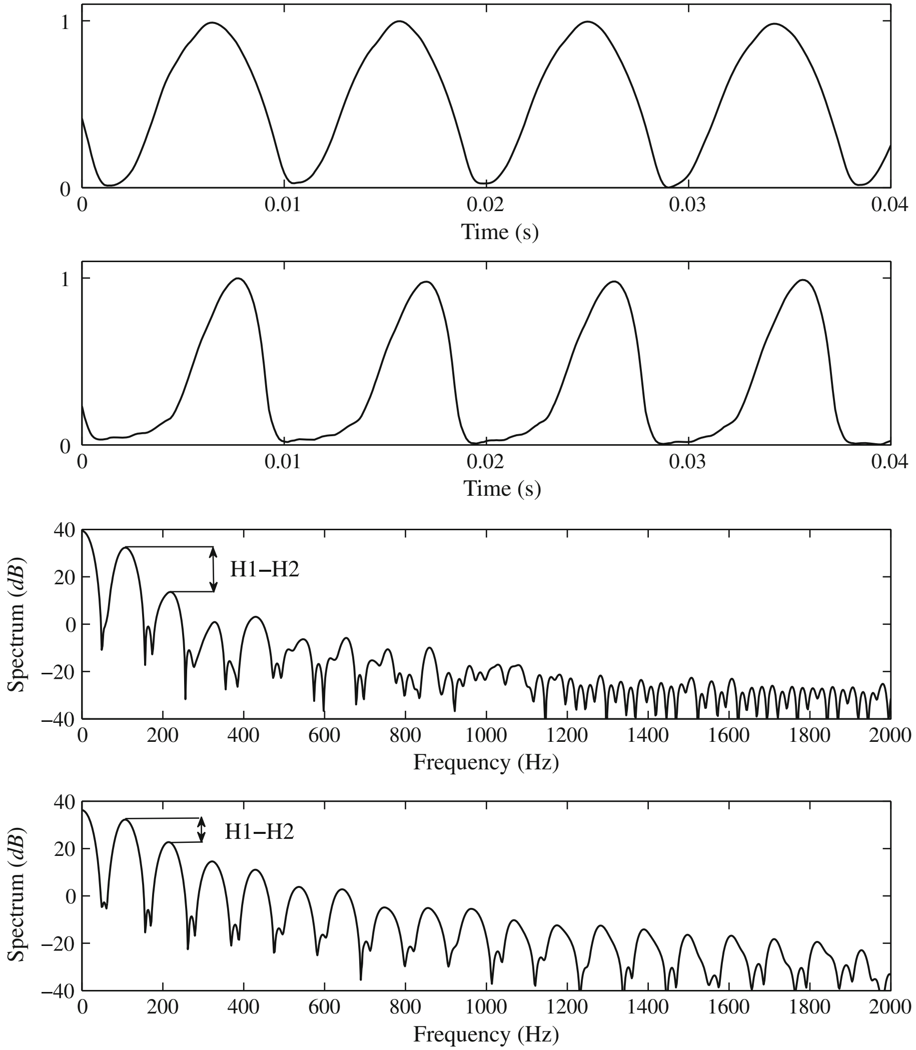

Figure 6. Glottal excitations estimated by inverse filtering the acoustical speech pressure waveforms pronounced by a male speaker. Time-domain waveforms of the glottal flow are shown in breathy (top panel) and pressed (second panel from top) phonation. Below these, voice source spectrum is shown for the breathy (third panel from top) and pressed (bottom panel) phonation. Different spectral decay of the two phonations is quantified by $H1-H2$: the value of $H1-H2$ is 18.4 dB and 9.6 dB in breathy and pressed phonation, respectively.

Finally, it is worth mentioning that there are also such measures of the source spectrum slope that do not utilize glottal inverse filtering *per se*. Instead these measures try to reflect the contribution of the glottal spectrum from spectral amplitudes, such as the level of the formants and harmonics, computed from the speech pressure signal (Hanson 1997; Hanson & Chuang 1999; Iseli *et al* 2007). These measures are discussed with a number of other spectral decay parameters in the study of Kreiman *et al* (2007). The reader is referred to the article by Kreiman *et al* (2007) to obtain more detailed information about relationships between the different frequency domain measures of the glottal source.

4. Applications of GIF

Since GIF provides a non-invasive method of computing the source of voiced speech from the microphone (or oral) recordings, it is understandable that the developed GIF methods have been applied for various purposes during the past four decades. Three main application[3] areas can be distinguished: (i) fundamental study of voice communication, including both speech and singing, (ii) medical applications, and (iii) speech technology applications. In this section, major studies where GIF techniques and parameterization methods have been applied are discussed.

In fundamental research of voice communication, there are plenty of topics, where GIF techniques have been used to understand how the characteristics of the glottal excitation vary in different voice production scenarios. One such topic is the *phonation type*, that is, the speaker's ability to change the fluctuation mode of the vocal folds to vary the glottal pulse from smooth to more abruptly changing (figure 6). This topic has been addressed by several authors using, for example, the Sondhi tube (e.g., Monsen & Engebretson 1977), the flow mask (e.g., Huffman 1987), the CP technique (e.g., Childers & Ahn 1995) and the Iterative Adaptive Inverse Filtering (IAIF) method (Alku & Vilkman 1996a) as the chosen method for estimating the glottal excitation. As changing the characteristics of the glottal source also affects the perceptual features of voice, GIF studies on phonation type are closely related to the following two large and general topics of speech science, the study of *voice quality* and the study of *vocal emotions*. Voice quality in this context is defined as the characteristic auditory colouring of a speaker's voice (Laver 1980) and it has been a topic of interest in numerous studies, where GIF techniques and voice source parameterization methods have been utilized (e.g., Price 1989; Gobl 1989; Childers & Lee 1991; Gobl & Ní Chasaide 1992; Campbell & Mokhtari 2003). In addition to these, there are many studies that have specifically focussed on the perceptual effects of the glottal source parameters by utilizing, for example, classical time-based parameters (Scherer *et al* 1998) or LF-parameters (van Dinther *et al* 2004; van Dinther *et al* 2005). Analysis of vocal emotions is an established topic of research that has been investigated especially in psychology for decades (Scherer 1986; Scherer 2003). A majority, of the studies are based on analysing parameters, such as sound pressure level (SPL) and F0, that depict the speech pressure signal, whereas the function of the glottal excitation has remained less studied mainly because of methodological difficulties. In contrast to this main trend of vocal emotion research, there are, however, certain GIF studies (e.g., Cummings & Clements 1995; Laukkanen *et al* 1996, Laukkanen *et al* 1997; Gobl & Ní Chasaide 2003b; Airas & Alku 2006; Waaramaa *et al* 2010) that have addressed the function of the glottal source in different emotional contexts. These studies have indicated that GIF, combined with, for example, the LF model or NAQ parameterization, gives valuable information related to the laryngeal production of emotional cues of speech. Another widely studied topic of human voice communication, where analysis of the glottal excitation plays an essential role, is *intensity regulation* (Titze 1994). Research in this topic addresses mechanisms underlying how speakers control the volume of their speech and how these mechanisms are affected by the characteristics of the glottal excitation. As intensity regulation is an essential feature in everyday speech communication, it is

[3] The word 'application' is interpreted in a wide sense in this study.

understandable that studying intensity regulation and its perceptual counterpart, vocal loudness, has raised interests among speech scientists over a period spanning several decades (Ladefoged & McKinney 1963; Isshiki 1964; Isshiki 1965; Gauffin and Sundberg 1989; Dromey *et al* 1992; Titze & Sundberg 1992; Stathopoulos & Sapienza 1993a; Stathopoulos & Sapienza 1993b; Sulter & Wit 1996; Alku *et al* 2006a; Seshadri & Yegnanarayana 2009). In one of the studies, the author specifically wishes to draw attention to the investigation conducted by Holmberg *et al* (1988), which involves a rich set of different voice source parameters and reports how they behave when speakers change the intensity of speech. GIF techniques have also been used in studying the *prosodic features* of speech. Examples thereof are studies conducted by Gobl (1988), Strik & Boves (1992), Ní Chasaide & Gobl (1993), Fant (1993; 1997), and Airas *et al* (2007). These investigations have addressed, for example, how the shape of the glottal excitation varies temporally within a phrase and how accentuation affects the voice source in connected speech. In addition to speech research, GIF methods have also been applied in the study of *singing* voices. In this area, especially the studies by Johan Sundberg are important. Analysis of the glottal excitation from singing has addressed, for example, the use of different registers (Björkner *et al* 2006), loudness control (Sundberg *et al* 1993; Sundberg *et al* 1999a; Sundberg *et al* 2005) and vibrato (Arroabarren & Carlosena 2004; Arroabarren & Carlosena 2006a). As F0 in singing is typically high, singing voices constitute sound material that can be considered highly challenging for GIF analysis. Therefore, some studies (e.g., Arroabarren & Carlosena 2006b) have criticized the use of GIF as an analysis method of singing voices. Finally, GIF methods have been applied in the fundamental study of speech to understand, for example, the effect of *ageing* (Higgins & Saxman 1991; Sapienza & Stathopoulos 1994; Sapienza & Dutka 1996; Hodge *et al* 2001) and *source-tract interaction* (Rothenberg 1985; Fant 1993; Childers & Wong 1994) in voice production. Studies on source tract interaction typically involve comparisons between the glottal area function and the volume velocity waveform.

Analysis and parameterization of the glottal volume velocity waveforms with GIF have also been conducted in medical sciences and related disciplines, such as in logopedics and voice therapy. In these areas, the most evident application for GIF is the *analysis of pathological voices*, which has been addressed in several studies (e.g., Koike & Markel 1975; Deller 1982; Hillman *et al* 1990; Howell & Williams 1988; Howell & Williams 1992). In addition to the analysis of pathological voices there are, also clinical studies where the glottal features have been used in the classification of speech which is emotionally disordered (Moore *et al* 2008). Although GIF methods have been utilized in the study of disordered speech signals as indicated by the references above, it is worth pointing out that the amount of studies published on normal speech is clearly much larger. This might be explained by the fact that the accuracy of GIF methods typically deteriorates when the speech signal becomes irregular and weak, which is characteristic for pathological voices. Hence, GIF might be more difficult to be applied for pathological sounds than for voices produced by healthy subjects. Nevertheless, there are reports, which indicate the GIF is an analysis technique that has proven useful in the clinical study of speech (Colton *et al* 1983; Hammarberg 2000). Furthermore, not all the studies in medical applications of GIF involve pathological voices but can address healthy voices which are in danger of becoming disordered due to, for example, *vocal loading* (e.g., Vilkman *et al* 1997; Lauri *et al* 1997, Lehto *et al* 2008). Vocal loading refers to a combination of prolonged voice use and additional loading factors such as background noise and air quality affecting the fundamental frequency, type and loudness of phonation and the vibratory patterns of the vocal folds (Vilkman 2004). Vocal loading is likely to affect especially people, such as teachers, who use their voice as the main tool in their work. These people, occupational voice users, are known to suffer from voice symptoms to varying extents (Fritzell 1996; Simberg *et al* 2000; Sala *et al* 2001). Owing to increasing number of employees working in such professions which call for extensive voice use (Titze *et al* 1997), the occupational voice care will become increasingly important in the future, which, in turn, will emphasize the role of the glottal flow parameterization techniques.

Methods to estimate the glottal excitation and to express the voice source parametrically have also been used in several areas of speech technology. The oldest and perhaps also the most evident area herein is *speech synthesis*. In this area, knowledge achieved via GIF has been used in the forms of artificial glottal source models and voice source control rules that have been utilized in several studies during the past three decades (e.g., Klatt 1987; Carlson *et al* 1989; Pinto *et al* 1989; de Veth *et al* 1990; Klatt & Klatt 1990; Carlson *et al* 1991; Karlsson 1991; Karlsson 1992; Fant 1993; Childers & Hu 1994; Childers 1995). In addition to these earlier studies, where the synthesis is typically rule-based and utilizes, for instance, the LF-model, GIF has also been used as a sound synthesis technique, in which the glottal source extracted from real speech is used to excite artificial vocal tract models (Holmes 1973; Matsui *et al* 1991; Alku *et al* 1999). These synthesis techniques enable generation of high-quality sounds, and the vocal tract characteristics can be controlled. Hence, the idea to excite synthetic speech by glottal waveforms extracted from natural utterances is well-suited especially for studies where brain activity evoked by speech is investigated and this principle has been taken advantage of in several studies (e.g., Näätänen *et al* 1997; Ceponiene *et al* 2003). It is also worth noting that in addition to earlier rule-based synthesis schemes, GIF methods have been recently combined with a state-of-the-art hidden Markov modelling (HMM)-based synthesis (Raitio *et al* 2011). Experiments reported by Raitio *et al* (2011) indicate that the physiologically oriented modelling approach of voice production, enabled by GIF, provides better synthesis quality than sound synthesis techniques that are typically used in current HMM-based synthesizers. As HMM-based synthesis is at present a research topic which attracts increasing interest from the international speech research community, it can be argued that the synthesis constitutes one of the most potential novel application areas for GIF methods. GIF-based methods have also been used in the sub-field of speech synthesis, *voice modification*, by processing either the pitch (e.g., Milenkovic 1993; Jiang & Murphy 2002) or other issues such as the phonation type (Childers 1995) or the speaking style (Cummings & Clements 1993). *Speaker identification* is the second main area of speech technology, where estimation of the glottal excitation with GIF has been applied. Although speech synthesis is used in implementing machine-to-man interfaces, speaker identification has applications in forensic speech research. An example of studies, where GIF methods have been utilized in speaker identification, is the investigations by Plumpe *et al* (1999) and Gudnason & Brookes (2008), who used the CP technique in the estimation of the glottal source. In the study by Plumpe *et al* (1999), the CP technique was utilized in a modified manner based on the analysis of formant frequency modulations. In addition, speaker identification has been studied with the IAIF technique by Lavner *et al* (2000) and Kinnunen & Alku (2009). Speaker identification has been also combined with *dialect identification* based on the GIF analysis of the glottal source (Yanguas *et al* 1999). The third main speech technology area, where GIF has been taken advantage of, is *speech coding* (Hedelin 1984; Hedelin 1986; Skoglund 1998). Although speech coding has been studied by utilizing GIF, it is, although, worth noting that the nature of these experiments is merely academic and none of them has resulted in applications which can be utilized in real communications systems. Speech coding is one of the areas of speech science that is characterized by the industrial standards and all the major speech coding methods used in real networks are based on different approaches, such as the algebraic code excited linear prediction ACELP method (ITU-T 1996).

5. Conclusions

Estimation of the source of voiced speech with GIF has been a goal of research for five decades. During this time, many GIF methods have been proposed and a wide range of glottal source parameterization techniques developed to express the estimated glottal excitation waveforms in numerical forms. Together these methods have enabled gathering basic science knowledge about the behaviour of the glottal excitation in different speech communication scenarios. In addition, the methods developed have been applied in several science disciplines, especially in medical voice research and speech technology.

GIF has proven to be an attractive analysis method of voice production because, *first*, the method is non-invasive (provided, the flow mask is not used). This is especially important in studies, such as the analysis of vocal emotions, where preserving natural voice production is essential. In addition, non-invasive analysis of voice production is a pre-requisite for studying occupational voice production, where the subject, such as teacher, must be able to conduct his or her normal duties without being limited by measurement equipment. *Second*, most GIF analyses can be conducted by using only one input, the speech pressure signal recorded in a free field by a microphone. Therefore, the analysis does not involve the use of special equipment, which makes recordings of the input data both inexpensive and fast. *Third*, it is at least in principle possible to implement both the inverse filtering and the parameterization stage in a completely automatic manner. *Fourth*, in comparison to some other widely used voice production analysis methods (e.g., EGG), the result yielded by GIF is a temporal signal, the glottal volume velocity waveform, which is an estimate of a real acoustical entity of the human voice production process. Owing to its direct relationship to the acoustical production of speech, estimates of glottal excitations computed by GIF can be practically used in speech technology applications, especially in speech synthesis.

Inverse filtering analysis of voice production also has several shortcomings. *First*, it has been reported in several studies (e.g., Holmes 1975; Wong *et al* 1979) that the method suffers from its sensitivity to the frequency characteristics of the input signal. In particular, if the estimated glottal excitation needs to be parameterized or visualized as a time-domain signal, which is the case in the majority of studies, the analysis calls for linear phase recordings. The reason for this can be explained by the general phase characteristics of the glottal pulse. It has been shown (Alku *et al* 2005) that the phase spectrum of a glottal pulse is almost a linear function of frequency except for the lowest frequency range. Hence, if this inherent phase behaviour is violated by non-linear recording equipment, such as the microphone, AD-card, or any pre-filtering before the inverse filtering analysis, the time-domain waveform of the output becomes distorted. This shortcoming of GIF analysis can be naturally avoided by using only such recording equipment, which is sufficiently phase linear over the frequency range that is beyond F0. Hence, this shortcoming does not restrict the use of GIF analysis unless the data to be inverse filtered involves pre-recorded speech samples that have been recorded with microphones of poor frequency characteristics or which have been, for example, transmitted through telephone systems. *Second*, GIF methods based on the processing of the speech pressure signal are unable to yield truthful amplitude-domain parameters from the glottal source. In particular, the lack of the DC flow might be regarded as a limitation if pathological voice production is to be analysed. This limitation, although, does not involve GIF analyses, where it is possible to use the flow mask. *Third*, the most severe shortcoming of GIF analysis is that of the accuracy in the estimation of the true glottal excitation. This problem is difficult not only due to methodological flaws in the inverse filtering algorithms but also because of difficulties in the evaluation of the accuracy. The ‘correct’ glottal source cannot be measured from the human larynx at least if natural voice production is to be preserved. Therefore, an estimate of the glottal pulseform computed using a GIF algorithm from natural speech cannot be compared with the 'correct' waveform to assess the algorithm's accuracy. As a solution, synthetic speech is typically used in testing the accuracy of the underlying GIF method. It is, although, worth noting that this approach is problematic if both the test material and the GIF method to be evaluated are based on the same assumptions about the functioning of the human speech production mechanism. Hence, using simple synthetic vowels generated by, for example, the source-tract model as test material is not in principle sufficient to assess the accuracy of a GIF method if the underlying method is also based on the same source-tract structure. As an alternative, some studies (Alku *et al* 2006b; Alku *et al* 2006c) have used synthetic speech produced by physical models of voice production, in which the test material is generated from the interaction of the self-sustained oscillation of the vocal folds with subglottal and supraglottal pressures. Hence, the test material is generated by physical law, rather than by a parametric waveform model. Evaluation of the GIF accuracy can also be approached by using multimodal analyses where information given by other voice production analysis methods, such as EGG, videostroboscopy or kymography, are used in parallel with acoustical inverse filtering analyses (e.g., Hertegård *et al* 1992; Hertegård & Gauffin 1995). In addition, assessment of GIF benefits from the use of a combined set of quality parameters as shown by the study of Moore and Torres (2008).

Although the accuracy of GIF methods is in principle difficult to quantify, several studies have concluded that the glottal source estimates tend to become especially unreliable in the analysis of certain voice types. The most typical example of this is represented by high-pitch speech, in particular if this occurs in parallel with a low F1. In this case, the sparse harmonic structure dominates the speech spectrum, which makes the extraction (both manual and automatic) of formant frequencies difficult and subject to estimation errors. Even small errors especially in the estimation of F1 will result in incomplete cancellation of the corresponding resonance of the vocal cavity, and the obtained estimate of the voice source will be distorted by a strong ripple component. High-pitch speech is especially sensitive for this kind of an artefact, because the corresponding 'true' glottal pulse typically has a short closed phase, which might completely disappear in the estimated pulse due to generation of the formant ripple. For the CP analysis, which is based on the estimation of the vocal tract model during the closed glottal phase, the increase in F0 implies that there are few data samples available to form an all-pole model for the vocal tract. Hence, results given by the CP analysis become increasingly more vulnerable to the estimation of the closed phase position when F0 increases. In addition to high-pitch speech, the estimation accuracy of GIF methods is limited for nasals and nasalized vowels because most of the methods use an all-pole structure in modelling of the vocal tract and are, therefore, unable to properly take into account the frequency effect of the mouth cavity in the production of nasal sounds. Finally, the accuracy of GIF methods is affected by the fact that most methods use such a digital filter in modelling of the vocal tract whose coefficients are fixed over a time period that spans several fundamental periods. This is a gross simplification of the real speech production mechanism, in which the frequency characteristics of the vocal cavity change during an individual glottal period due to coupling of the subglottal system during the open phase of the glottal cycle and the uncoupling of the same system when the vocal folds are closed.

All in all, the estimation of the glottal excitation with GIF has both strong benefits, (especially in terms of the non-invasive nature of the analysis) and promising novel applications (such as HMM-based speech synthesis), but also severe shortcomings (in particular, that of accuracy for high-pitch speech). There is undoubtedly scope for new research and innovations in this study area. Combining novel computation inversion mathematics which enable involving *a priori* information about the position of the vocal tract resonances could be a potential avenue, which might shed light on the methodological shortcomings referred to earlier. In addition, expanding the GIF analysis from the analysis of sustained vowels, which has been a major trend in the study area for decades, into the analysis of continuous speech would undoubtedly be worthwhile to elucidate the behaviour of the glottal function in real speech communication. It is worth emphasising that having a GIF method capable of estimating the *entire* glottal source from continuous speech might be a distant dream. However, expanding the GIF analysis to involve continuous speech becomes feasible even with existing techniques if one, for example, analyses the GIF only for those sections that are most reliable to be analysed with GIF (i.e., long vowels with high F1). Automatic analysis could be enabled by combing speech recognition techniques with the proposed GIF analysis.

This study has been supported by the Academy of Finland (Lastu Programme, project no. 135003).

References

Airas M 2008 TKK Aparat: An environment for voice inverse filtering and parameterization, *Logoped. Phoniatr. Vocol.* 33: 49–64

Airas M, Alku P 2006 Emotions in vowel segments of continuous speech, *Phonetica* 63: 26–46

Airas M, Alku P, Vainio M 2007 Laryngeal voice quality changes in expression of prominence in continuous speech. *Proc. 5th Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA'07)*, 135–138

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

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

Alku P, Airas M, Björkner E, Sundberg J 2006a An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity, *J. Acoust. Soc. Am.* 120(1): 1052–1062

Alku P, Airas M, Bäckström T, Pulakka H 2005 Using group delay function to assess glottal flows estimated by inverse filtering, *Electron. Lett.* 41(9): 562–563

Alku P, Bäckström T, Vilkman E 2002 Normalized amplitude quotient for parameterization of the glottal flow, *J. Acoust. Soc. Am.* 112(1): 701–710

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

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

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

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

Alku P, Vilkman E 1994 Estimation of the glottal pulseform based on discrete all-pole modeling. *Proc. Int. Conference on Spoken Language Processing (ICSLP)* 3: 1619–1622

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

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

Ananthapadmanabha T, Yegnanarayana B 1975 Epoch extraction of voiced speech, *IEEE Trans. Acoust. Speech Signal Process.* 23(6): 562–570

Arroabarren I, Carlosena A 2003 Glottal spectrum based inverse filtering. *Proc. Eurospeech*, 57–60

Arroabarren I, Carlosena A 2004 Vibrato in singing voice: The link between source-filter and sinusoidal models, *EURASIP J. Appl. Signal Process.* 7: 1007–1020

Arroabarren I, Carlosena A 2006a Effect of the glottal source and the vocal tract on the partials amplitude of vibrato in male voices, *J. Acoust. Soc. Am.* 119(4): 2483–2497

Arroabarren I, Carlosena A 2006b Inverse filtering in singing voice: a critical analysis, *IEEE Trans. Audio Speech Lang. Process.* 14(4): 1422–1431

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

Avanzini F, Alku P, Karjalainen M 2001 One-delayed-mass model for efficient synthesis of glottal flow. *Proc. Eurospeech*, 51–54

Björkner E, Sundberg J, Cleveland T, Stone E 2006 Voice source differences between registers in female musical theater singers, *J. Voice* 20(1): 187–197

Boves L, Cranen B 1982 Evaluation of glottal inverse filtering by means of physiological registrations. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 3: 1988–1991

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

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

Bäckström T, Airas M, Lehto L, Alku P 2005 Objective quality measures for glottal inverse filtering of speech pressure signals. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 1: 897–900

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

Campbell N, Mokhtari P 2003 Voice quality: the 4th prosodic dimension. *Proc. of the 15th Int. Congress of Phonetic Sciences* 3: 2417–2420

Carlson R, Fant G, Gobl C, Granström B, Karlsson I, Lin Q-G 1989 Voice source rules for text-to-speech synthesis. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 1: 223–226

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

Carr P, Trill D 1964 Long-term larynx-excitation spectra, *J. Acoust. Soc. Am.* 36(11): 2033–2040

Catford J C 1977 *Fundamental problems in phonetics* (Edinburgh, UK: Edinburgh University Press)

Ceponiene R, Lepistö T, Shestakova A, Vanhala R, Alku P, Näätänen R, Yaguchi K 2003 Speech-sound selective auditory impairment in autism: can perceive but do not attend, *Proc. Natl. Acad. Sci. U.S.A.* 100(9): 5567–5572

Childers D 1995 Glottal source modeling for voice conversion, *Speech Commun.* 16: 127–138

Childers D, Ahn C 1995 Modeling the glottal volume-velocity waveform for three voice types, *J. Acoust. Soc. Am.* 97(1): 505–519

Childers D, Hu H 1994 Speech synthesis by glottal excited linear prediction, *J. Acoust. Soc. Am.* 96(4): 2026–2036

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

Childers D, Principe J, Ting Y 1995 Adaptive WRLS-VFF for speech analysis, *IEEE Trans. Speech Audio Process.* 3(2): 209–213

Childers D, Wong C-F 1994 Measuring and modeling vocal source-tract interaction, *IEEE Trans. Biomed. Eng.* 41(7): 663–671

Colton R, Brewer D, Rothenberg M 1983 Evaluating vocal fold function, *J. Otolaryngology* 12(5): 291–294

Cranen B, Boves L 1988 On the measurement of glottal flow, *J. Acoust. Soc. Am.* 84(2): 888–900

Cummings K, Clements M 1993 Application of the analysis of glottal excitation of stressed speech to speaking style modification. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 2: 207–210

Cummings K, Clements M 1995 Analysis of the glottal excitation of emotionally styled and stressed speech, *J. Acoust. Soc. Am.* 98: 88–98

Dalsgaard P, Pedersen C, Andersen O, Yegnanarayana B 2008 Using zeros of the z-transform in the analysis of speech signals. *CD Proc. of the ISCA Tutorial and Research Workshop, Speech Analysis and Processing for Knowledge Discovery*

Deller J Jr 1982 Evaluation of laryngeal dysfunction based on features of an accurate estimate of the glottal waveform. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 2: 759–762

Deng H, Ward R, Beddoes M, Hodgson M 2006 A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds, *IEEE Trans. Audio Speech Lang. Process.* 14(1): 445–455

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

Ding W, Campbell N, Higuchi N, Kasuya H 1997 Fast and robust joint estimation of vocal tract and voice source parameters. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 2: 1291–1294

Drioli C 2005 A flow waveform-matched low-dimensional glottal model based on physical knowledge, *J. Acoust. Soc. Am.* 117(5): 3184–3195

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

Drugman T, Bozkurt B, Dutoit T 2009 Complex cepstrum-based decomposition of speech for glottal source estimation. *CD Proc. Interspeech*, 116–119

Edwards J, Angus J 1996 Using phase-plane plots to assess glottal inverse filtering, *Electron. Lett.* 32(2): 192–193

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

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

Fant G 1961 A new anti-resonance circuit for inverse filtering, *Speech Transmission Laboratory Quarterly Progress and Status Report* 2(1):1–6

Fant G 1970 *Acoustic theory of speech production* (The Hague, Netherlands: Mouton)

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

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

Fant G 1997 The voice source in connected speech, *Speech Commun.* 22: 125–139

Fant G, Kruckenberg A, Liljencrants J, Båvegård M 1994 Voice source parameters in continuous speech. Transformation of LF-parameters. *Proc. Int. Conference on Spoken Language Processing (ICSLP)* 3: 1451–1454

Fant G, Liljencrants J, Lin Q 1985 A four-parameter model of glottal flow, *Speech Transmission Laboratory Quarterly Progress and Status Report* 26(4): 1–13

Fant G, Lin Q 1988 Frequency domain interpretation and derivation of glottal flow parameters, *Speech Transmission Laboratory Quarterly Progress and Status Report* 29(2-3): 1–21

Fant G, Sonesson B 1962 Indirect studies of glottal cycles by synchronous inverse filtering and photo-electrical glottography, *Speech Transmission Laboratory Quarterly Progress and Status Report* 3(4): 1–3

Flanagan J 1972 *Speech analysis, synthesis and perception* (New York, NY: Springer Verlag)

Fritzell B 1992 Inverse filtering, *J. Voice* 6(1): 111–114

Fritzell B 1996 Voice disorders and occupations, *Logoped. Phoniatr. Vocol.* 21: 7–12

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

Fröhlich M, Michaelis D, Strube H 2001 SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals, *J. Acoust. Soc. Am.* 110(1): 479–488

Fu Q, Murphy P 2003 Adaptive inverse filtering for high accuracy estimation of the glottal source. *Proc. of the ISCA Tutorial and Research Workshop, Non-Linear Speech Processing*, paper 018

Fu Q, Murphy P 2006 Robust glottal source estimation based on joint source-filter model optimization, *IEEE Trans. Audio Speech Lang. Process.* 14(1): 492–501

Fujisaki H, Ljungqvist M 1986 Proposal and evaluation of models for the glottal source waveform. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 3: 1605–1608

Fujisaki H, Ljungqvist M 1987 Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 2: 637–640

Gauffin J, Sundberg J 1989 Spectral correlates of glottal voice source waveform characteristics, *J. Speech Hear. Res.* 32: 556–565

Gobl C 1988 Voice source dynamics in connected speech, *Speech Transmission Laboratory Quarterly Progress and Status Report* 29(1): 123–159

Gobl C 1989 A preliminary study of acoustic voice quality correlates, *Speech Transmission Laboratory Quarterly Progress and Status Report* 30(4): 9–22

Gobl C 2003 The voice source in speech communication. Production and perception experiments involving inverse filtering and synthesis. Doctoral thesis. Royal Institute of Technology, Stockholm, Sweden.

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

Gobl C, Ní Chasaide A 2003a Amplitude-based source parameters for measuring voice quality. *CD Proc. of the ISCA Tutorial and Research Workshop, Voice Quality: Functions, Analysis, and Synthesis*

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

Granqvist S, Hertegård S, Larsson H, Sundberg J 2003 Simultaneous analysis of vocal fold vibration and transglottal airflow: exploring a new experimental set-up, *J. Voice* 17(2): 319–330

Gudnason J, Brookes M 2008 Voice source cepstrum coefficients for speaker identification. *CD Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 4821–4824

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

Hammarberg B 2000 Voice research and clinical needs, *Folia Phoniatr. Logo.* 52: 93–102

Hanson H 1997 Glottal characteristics of female speakers: Acoustic correlates, *J. Acoust. Soc. Am.* 101(1): 466–481

Hanson H, Chuang E 1999 Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *J. Acoust. Soc. Am.* 106(1): 1064–1077

Hedelin P 1984 A glottal LPC-vocoder. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 1: 1.6.1–1.6.4

Hedelin P 1986 High quality glottal LPC-vocoding. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 1: 465–468

Hertegård S, Gauffin J 1992 Acoustic properties of the Rothenberg mask, *Speech Transmission Laboratory Quarterly Progress and Status Report* 33(2-3): 9–18

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

Hertegård S, Gauffin J, Karlsson I 1992 Physiological correlates of the inverse filtered flow waveform, *J. Voice* 6: 224–234

Higgins M, Saxman J 1991 A comparison of selected phonatory behaviors of healthy aged and young adults, *J. Speech Hear. Res.* 34: 1000–1010

Hillman R, Holmberg E, Perkell J, Walsh M, Vaughan C 1990 Phonatory function associated with hyperfunctionally related vocal fold lesions, *J. Voice* 4(1): 52–64

Hillman R, Weinberg B 1981 Estimation of glottal volume velocity waveform properties: A review and study of some methodological assumptions. In *Speech and Language: Advances in Basic Research and Practice*, N Lass (ed.), Academic: New York, 411–473

Hirano M 1981 *Clinical examination of voice* (New York, NY: Springer Verlag)

Hodge S, Colton R, Kelley R 2001 Vocal intensity characteristics in normal and elderly speakers, *J. Voice* 15(4): 503–511

Holmberg E, Hillman R, Perkell J 1988 Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice, *J. Acoust. Soc. Am.* 84: 511–529

Holmes J 1963 An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter. *Proc. of the Speech Communication Seminar*, Paper B4

Holmes J 1973 The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer, *IEEE Trans. Audio Electroacoust.* 21(2): 298–305

Holmes J 1975 Low-frequency phase distortion of speech recordings, *J. Acoust. Soc. Am.* 58(2): 747–749

Holmes J 1976 Formant excitation before and after glottal closure. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 1: 39–42

Howell P, Williams M 1988 The contribution of the excitatory source to the perception of neutral vowels in stuttered speech, *J. Acoust. Soc. Am.* 84(1): 80–89

Howell P, Williams M 1992 Acoustic analysis and perception of vowels in children's and teenagers' stuttered speech, *J. Acoust. Soc. Am.* 91(2): 1697–1706

Huffman M 1987 Measures of phonation type in Hmong, *J. Acoust. Soc. Am.* 81(1): 495–504

Hunt M 1987 Studies of glottal excitation using inverse filtering and an electroglottograph. *Proc. of the 11th Int. Congress of Phonetic Sciences* 3: 23–26

Hunt M, Bridle J, Holmes J 1978 Interactive digital inverse filtering and its relation to linear prediction methods. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 3: 15–18

Isaksson A, Millnert M 1989 Inverse glottal filtering using a parameterized input model, *Signal Process.* 18: 435–445

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

Isshiki N 1964 Regulatory mechanism of voice intensity variation, *J. Speech Hear. Res.* 7: 17–29

Isshiki N 1965 Vocal intensity and air flow rate, *Folia Phoniatr.* 17: 92–104

Isshiki N 1981 Vocal efficiency index. In K.N. Stevens and M. Hirano (eds), *Vocal fold physiology*, Tokyo: University of Tokyo Press, 193–203

ITU-T 1996 Recommendation G.729. Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)

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

Javkin H, Antonanzas-Barroso N, Maddieson I 1987 Digital inverse filtering for linguistic research, *J. Speech Hear. Res.* 30: 122–129

Jiang Y, Murphy P 2002 Production based pitch modification of voiced speech. *CD Proc. Int. Conference on Spoken Language Processing (ICSLP)*, 2073–2076

Kaburagi T, Kawai K 2003 Analysis of voice source characteristics using a constrained polynomial model. *CD Proc. Eurospeech*, 461–464

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

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

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

Kinnunen T, Alku P 2009 On separating glottal source and vocal tract information in telephony speaker verification. *CD Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 4545–4548

Klatt D 1987 Review of text-to-speech conversion for English, *J. Acoust. Soc. Am.* 82(2): 737–793

Klatt D, Klatt L 1990 Analysis, synthesis, and perception of voice quality variations among female and male talkers, *J. Acoust. Soc. Am.* 87(1): 820–857

Koike Y, Markel J 1975 Application of inverse filtering for detecting laryngeal pathology, *Ann. Otoal. Rhinol. Laryngol.* 84: 117–124

Kreiman J, Gerratt B, Antonanzas-Barroso N 2007 Measures of the glottal source spectrum, *J. Speech Lang. Hear. Res.* 50: 595–610

Krishnamurthy A, Childers D 1986 Two-channel speech analysis, *IEEE Trans. Acoust. Speech Signal Process.* 34: 730–743

Ladefoged P, McKinney NP 1963 Loudness, sound pressure, and subglottal pressure in speech, *J. Acoust. Soc. Am.* 35: 454–460

Larar J, Alsaka Y, Childers D 1985 Variability in closed phase analysis of speech. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 3: 1089–1092

Laukkanen A-M, Vilkman E, Alku P, Oksanen H 1996 Physical variations related to stress and emotional state: A preliminary study, *J. Phonetics* 24: 313–335

Laukkanen A-M, Vilkman E, Alku P, Oksanen H 1997 On the perception of emotions in speech: The role of voice quality, *Logoped. Phoniatr. Vocol.* 22: 157–168

Lauri E-R, Alku P, Vilkman E, Sala S, Sihvo M 1997 Effects of prolonged oral reading on time-based glottal flow waveform parameters with special reference to gender differences, *Folia Phoniatr. Logo.* 49: 234–246

Laver J 1980 *The phonetic description of voice quality* (Cambridge, UK: Cambridge University Press)

Lavner Y, Gath I, Rosenhouse J 2000 The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels, *Speech Commun.* 30: 9–26

Lecluse F, Brocaar M, Verschuure J 1975 The electroglottography and its relation to glottal activity, *Folia Phoniatr.* 17: 215–224

Lehto L, Laaksonen L, Vilkman E, Alku P 2008 Changes in objective acoustic measurements and subjective voice complaints in call-center customer-service advisors during one working day, *J. Voice* 22(1): 164–177

Lindqvist-Gauffin J 1963 Inverse filtering equipment, *Speech Transmission Laboratory Quarterly Progress and Status Report* 4(1): 13

Lindqvist-Gauffin J 1964 Inverse filtering. Instrumentation and techniques, *Speech Transmission Laboratory Quarterly Progress and Status Report* 5(4): 1–4

Lindqvist-Gauffin J 1965 Studies of the voice source by means of inverse filtering, *Speech Transmission Laboratory Quarterly Progress and Status Report* 6(1): 8–13

Lu H, Smith J 1999 Joint estimation of vocal tract filter and glottal source waveform via convex optimization. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 79–82

Matausek M, Batalov V 1980 A new approach to the determination of the glottal waveform, *IEEE Trans. Acoust. Speech Signal Process.* 28(6): 616–622

Mathews M, Miller J, David E 1961 Pitch synchronous analysis of voiced speech, *J. Acoust. Soc. Am.* 33(1): 179–186

Matsui K, Pearson S, Hata K, Kamai T 1991 Improving naturalness in text-to-speech synthesis using natural glottal source. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 2: 769–772

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

Milenkovic P 1993 Voice source model for continuous control of pitch period, *J. Acoust. Soc. Am.* 93(1): 1087–1096

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

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

Moore E, Clements M 2004 Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 1: 101–104

Moore E, Clements M, Peifer J, Weisser L 2008 Critical analysis of the impact of glottal features in the classification of clinical depression in speech, *IEEE Trans. Biomed. Eng.* 55(1): 96–107

Moore E, Torres J 2008 A performance assessment of objective measures for evaluating the quality of glottal waveform estimates, *Speech Commun.* 50(1): 56–66

Murphy P 1999 Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis, *J. Acoust. Soc. Am.* 105(5): 2866–2881

Murty K, Yegnanarayana B 2008 Epoch extraction from speech signals, *IEEE Trans. Audio Speech Lang. Process.* 16(8): 1602–1613

Nakatsui M, Suzuki J 1970 Method of observation of glottal-source wave using digital inverse filtering in time domain, *J. Acoust. Soc. Am.* 47(1): 664–665

Ní Chasaide A, Gobl C 1993 Contextual variation of the vowel voice source as a function of adjacent consonants, *Lang. Speech* 36: 303–330

Näätänen R, Lehtokoski A, Lennes M, Cheour M, Huotilainen M, Iivonen A, Vainio M, Alku P, Ilmoniemi R, Luuk A, Allik J, Sinkkonen A, Alho K 1997 Language-specific phoneme representations revealed by electric and magnetic brain responses, *Nature* 385: 432–434

Oppenheim A, Schafer R 1968 Homomorphic analysis of speech, *IEEE Trans. Audio Electroacoust.* 16(1): 221–226

Perez J, Bonafonte A 2009 Towards robust glottal source modeling. *Proc. Interspeech*, 68–71

Pinto N, Childers D, Lalwani A 1989 Formant speech synthesis: Improving production quality, *IEEE Trans. Acoust. Speech Signal Process.* 37(12): 1870–1887

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

Price P 1989 Male and female voice source characteristics: Inverse filtering results, *Speech Commun.* 8: 261–277

Rabiner L, Schafer R 1978 *Digital processing of speech signals* (Englewood Cliffs, NY: Prentice-Hall)

Raitio T, Suni A, Yamagishi Y, Pulakka H, Nurminen J, Vainio M, Alku P 2011 HMM-based speech synthesis utilizing glottal inverse filtering, *IEEE Trans. Audio Speech Lang. Process.* 19(1): 153–165

Rao KS, Yegnanarayana B 2006 Prosody modification using instants of significant excitation, *IEEE Trans. Audio Speech Lang. Process.* 14(2): 972–980

Riegelsberger E, Krishnamurthy A 1993 Glottal source estimation: Methods of applying the LF-model to inverse filtering. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 2: 542–545

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

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

Rothenberg M 1977 Measurement of airflow in speech, *J. Speech Hear. Res.* 20: 155–176

Rothenberg M 1985 Source-tract acoustic interaction in breathy voice. *Vocal fold physiology: biomechanics, acoustics and phonatory control*. I Titze and R Scherer (eds), Denver, Colorado: The Denver Center for the Performing Arts, 465–482

Rothenberg M, Zahorian S 1977 Nonlinear inverse filtering technique for estimating the glottal-area waveform, *J. Acoust. Soc. Am.* 61(4): 1063–1071

Sala E, Laine A, Simberg S, Pentti J, Suonpää J 2001 The prevalence of voice disorders among day care center teachers compared with nurses: A questionnaire and clinical study, *J. Voice* 15(2): 413–423

Sapienza C, Dutka J 1996 Glottal airflow characteristics of women's voice production along an aging continuum, *J. Speech Hear. Res.* 39: 322–328

Sapienza C, Stathopoulos E 1994 Comparison of maximum flow declination rate: Children versus adults, *J. Voice* 8: 240–247

Sapienza C, Stathopoulos E, Dromey C 1998 Approximations of open quotient and speed quotient from glottal airflow and EGG waveforms: Effects of measurement criteria and sound pressure level, *J. Voice* 12(1): 31–43

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

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

Scherer R, Arehart K, Guo C, Milstein C, Horii Y 1998 Just noticeable differences for glottal flow waveform characteristics, *J. Voice* 12(1): 21–30

Schnell K, Lacroix A 2007 Time-varying pre-emphasis and inverse filtering of speech. *CD Proc. Interspeech*, 530–533

Seshadri G, Yegnanarayana B 2009 Perceived loudness of speech based on the characteristics of glottal excitation source, *J. Acoust. Soc. Am.* 126(4): 2061–2071

Shapira Y, Gath I 1998 A geometrical fuzzy clustering-based solution to glottal wave estimation, *J. Acoust. Soc. Am.* 104(5): 3070–3079

Shiga Y, King S 2003 Estimation of voice source and vocal tract characteristics based on multi-frame analysis. *CD Proc. Eurospeech*, 1749–1752

Simberg S, Laine A, Sala E, Rönnemaa A-M 2000. Prevalence of voice teachers among future teachers, *J. Voice* 14(2): 231–235

Skoglund J 1998 Analysis and quantization of glottal pulse shapes, *Speech Commun.* 24: 133–152

Smits R, Yegnanarayana B 1995 Determination of instants of significant excitation in speech using group delay function, *IEEE Trans. Speech Audio Process.* 3(5): 325–333

Sondhi M 1975 Measurement of the glottal waveform *J. Acoust. Soc. Am.* 57(1): 228–232

Sonesson B 1959 A method for studying the vibratory movements of the vocal cords. A preliminary report, *J. Laryngol.* 73: 732–737

Stathopoulos E, Sapienza C 1993a Respiratory and laryngeal function of women and men during vocal intensity variation, *J. Speech Hear. Res.* 36: 64–75

Stathopoulos E, Sapienza C 1993b Respiratory and laryngeal measures of children during vocal intensity variation, *J. Acoust. Soc. Am.* 94: 2531–2543

Strik H, Boves L 1992 On the relation between voice source parameters and prosodic features in connected speech, *Speech Commun.* 11: 167–174

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

Sturmel N, D'Alessandro C, Doval B 2007 A comparative evaluation of the zeros of Z transform representation for voice source estimation. *CD Proc. Interspeech*, 558–561

Sulter A, Wit H 1996 Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age, *J. Acoust. Soc. Am.* 100(5): 3360–3373

Sundberg J, Andersson M, Hultqvist C 1999a Effects of subglottal pressure on professional baritone singers' voice sources, *J. Acoust. Soc. Am.* 105(2): 1965–1971

Sundberg J, Cleveland T, Stone R, Iwarsson J 1999b Voice source characteristics in six premier country singers, *J. Voice* 13(1): 168–183

Sundberg J, Fahlstedt E, Morell A 2005 Effects on the glottal voice source of vocal loudness variation in untrained female and male voices, *J. Acoust. Soc. Am.* 117(2): 879–885

Sundberg J, Titze I, Scherer R 1993 Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source, *J. Voice* 7: 15–29

Švec J, Schutte H 1996 Videokymography: high-speed line scanning of vocal fold vibration, *J. Voice* 10: 201–205

Södersten M, Håkansson A, Hammarberg B 1999 Comparison between automatic and manual inverse filtering procedures for healthy female voices, *Logoped. Phoniatr. Vocol.* 24: 26–38

Thomson M 1992 A new method for determining the vocal tract transfer function and its excitation from voiced speech. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 2: 37–40

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

Titze I 1994 *Principles of voice production* (Englewood Cliffs, NJ: Prentice-Hall)

Titze I, Lemke J, Montequin D 1997 Populations in the US workforce who rely on voice as a primary tool of trade: A preliminary report, *J. Voice* 11: 245–259.

Titze I, Story B, Burnett G, Holzrichter J, Ng L, Lea W 2000 Comparison between electroglottography and electromagnetic glottography, *J. Acoust. Soc. Am.* 107: 581–588

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

van Dinther R, Kohlrausch A, Veldhuis R 2004 A method for analysing the perceptual relevance of glottal-pulse parameter variations, *Speech Commun.* 42: 175–189 van Dinther R, Veldhuis R, Kohlrausch A 2005 Perceptual aspects of glottal-pulse parameter variations, *Speech Commun.* 46: 95–112

Veeneman D, BeMent S 1985 Automatic glottal inverse filtering from speech and electroglottographic signals, *IEEE Trans. Acoust. Speech Signal Process.* 33: 369–377

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

Vilkman E 2004 Occupational safety and health aspects of voice and speech professions, *Folia Phoniatr. Logo.* 56: 220–253

Vilkman E, Lauri E-R, Alku P, Sala E, Sihvo M 1997 Loading changes in time based parameters of glottal flow waveforms in different ergonomic conditions, *Folia Phoniatr. Logo.* 49: 247–263

Waaramaa T, Laukkanen A-M, Airas M, Alku P 2010 Perception of emotional valences from vowel segments of continuous speech, *J. Voice* 24(1): 30–38

Walker J, Murphy P 2005 Advanced methods for glottal wave extraction. In *Nonlinear analyses and algorithms for speech processing*, M Faundez-Zanuy *et al* (eds), Springer Verlag: Berlin, 139–149

Wong D, Markel J, Gray A 1979 Least squares glottal inverse filtering from the acoustic speech waveform, *IEEE Trans. Acoust. Speech Signal Process.* 27: 350–355

Yanguas L, Quatieri T, Goodman F 1999 Implications of glottal source for speaker and dialect identification. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 2: 813–816

Yegnanarayana B, Murty K 2009 Event-based instantaneous fundamental frequency estimation from speech signals, *IEEE Trans. Audio Speech Lang. Process.* 17(4): 614–624

Yegnanarayana B, Veldhuis N 1998 Extraction of vocal-tract system characteristics from speech signals, *IEEE Trans. Speech Audio Process.* 6: 313–327