Glottal Source Processing: from Analysis to Applications

Thomas Drugman[a,*], Paavo Alku[b], Abeer Alwan[c], Bayya Yegnanarayana[d]

[a] *TCTS Lab, University of Mons, Belgium*
[b] *Department of Signal Processing and Acoustics, Aalto University, Finland*
[c] *Speech Processing and Auditory Perception Laboratory, University of California Los Angeles, USA*
[d] *Speech and Vision Laboratory, International Institute of Information Technology, Hyderabad, India*

Abstract

The great majority of current voice technology applications rely on acoustic features, such as the widely used MFCC or LP parameters, which characterize the vocal tract response. Nonetheless, the major source of excitation, namely the glottal flow, is expected to convey useful complementary information. The glottal flow is the airflow passing through the vocal folds at the glottis. Unfortunately, glottal flow analysis from speech recordings requires specific and complex processing operations, which explains why it has been generally avoided. This paper gives a comprehensive overview of techniques for glottal source processing. Starting from analysis tools for pitch tracking, detection of glottal closure instant, estimation and modeling of glottal flow, this paper discusses how these tools and techniques might be properly integrated in various voice technology applications.

Keywords: Speech processing, Speech analysis, Glottal source, Glottal Flow, Excitation signal, Residual signal

[*] Corresponding author. Tel. +3265374749.
Email addresses: thomas.drugman@umons.ac.be (Thomas Drugman), paavo.alku@aalto.fi (Paavo Alku), alwan@ee.ucla.edu (Abeer Alwan), yegna@iiit.ac.in (Bayya Yegnanarayana)

1. Introduction

During the mechanism of phonation, an airflow evicted from the lungs, and passing through the trachea is modulated by vibrating the vocal folds, at the glottis (Quatieri, 2002). Speech is a result of filtering this so-called glottal flow by the vocal tract cavities, and converting the resulting velocity flow into pressure at the lips (Quatieri, 2002). In many speech processing applications, it is important to separate the contributions from the source at the glottis and the resonances due to vocal tract. Achieving such a *source-filter deconvolution* could lead to a distinct characterization and modeling of these two components, as well as to a better understanding of the human phonation process. Glottal source processing refers to the methods targeting the estimation, modeling and characterization of the glottal source, as well as the integration of related information within various voice technology applications. In this paper, we limit our scope to the methods which perform an analysis directly from the speech waveform, as captured by a microphone.

Glottal source processing is necessary for the study of glottal-based vocal effects, which can be segmental (as for vocal fry), or be controlled by speakers on a separate supra-segmental layer (as in the case for the voice quality modifications), and whose dynamics are very different from that of the vocal tract contribution. As will be further emphasized in Section 5, information about the glottal production can be effectively used in several applications such as speech synthesis, expressive speech processing, speaker recognition, and voice-based biomedical engineering.

The techniques estimating and parameterizing the vocal tract response are well established and have reached a certain maturity using, for example, the relatively widely-used Mel Frequency Cepstral Coefficients (MFCC) or Linear Predicition (LP) parameters. But it is not the case for glottal source processing methods, although this topic is gaining attention, and some advances have recently been made. One reason for this is that algorithms for glottal source processing have typically to be synchronous with the glottal cycles, unlike the usual speech analysis and feature extraction methods.

Figure 1 illustrates the workflow of a complete glottal source processing system, starting from the speech signal through to the integration of glottal information in a voice technology application. In contrast to a conventional feature extraction method (with fixed frame size and frame shift as in MFCC or LP analysis), the characterization of the glottal flow is more involved. The glottal source process can be divided into three main steps: *i)* synchronization with the glottal cycles, *ii)* estimating and parameterizing the glottal source, and *iii)* incorporating the glottal information in the target application. In the first step, the frequency of vibration of the vocal folds (called instantaneous fundamental frequency (F_0)), the polarity of the speech signal and the instants of significant excitation in each glottal cycle (called Glottal Closure Instants (GCIs)) are determined. This information is necessary for analysis requiring a glottal-synchronous approach. In the second step, the glottal flow is estimated from the speech signal. This is a difficult blind separation problem, since neither the vocal tract response nor the glottal contribution are observable. The glottal flow estimate is then parameterized by extracting relevant features. In the last step, the resulting glottal information is used to improve the performance of a given voice technology application. The integration of the glottal information depends upon the target application.

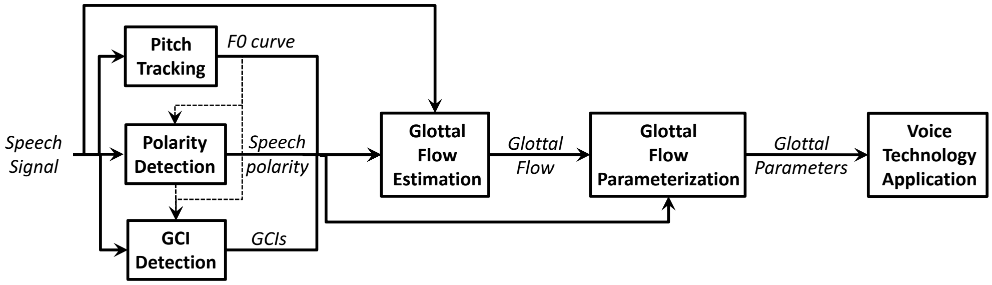

Figure 1: *Workflow of a complete glottal source processing system. Dashed arrows represent optional data flows.*

The objective of this paper is to present a comprehensive overview of the current state of the art in glottal source processing. Contrary to existing reviews, which have focused on a specific topic of glottal source analysis such as GCI detection (Drugman et al., 2012c) or glottal source estimation and parameterization (Walker and Murphy, 2007; Alku, 2011), our goal here is to cover all issues in the workflow depicted in Figure 1. The paper is structured as follows. Section 2 introduces the physiology of glottal production by describing briefly the organs that contribute to the glottal flow. Section 3 discusses the important features of glottal vibration, namely, the perception of pitch and its variation due to instantaneous fundamental frequency (F_0), speech polarity detection and GCI determination. This section also discusses the issues involved in extracting these features from the speech signal. Section

4 deals with estimation of the glottal source from the speech signal. Section 4 also considers parameterization of the extracted glottal source information both in the time-domain as well as in the frequency-domain. The significance of glottal source processing can be appreciated in the context of different applications. Hence, the applicability of glottal source processing is illustrated in four application scenarios in Section 5: speech synthesis, expressive speech analysis, speaker recognition, and biomedical applications. Finally, Section 6 provides a summary and conclusions in the form of identifying challenges in glottal source processing from speech collected in practical environments.

2. Physiology of glottal production

The human speech production system allows a speaker to produce a vast range of sounds. The system consists of many organs cooperating in the phonation process, which can be generally categorized into three groups: the lungs, the larynx, and the vocal tract. From a physiological point of view, the airflow from the lungs is pushed through the larynx, where the airflow is modulated by the vibration of the vocal folds (also known as the vocal cords). The vocal fold vibration converts the airflow to acoustic pulses and provides an excitation signal to the vocal tract. The vocal tract consists of the oral, nasal, and pharyngeal resonant cavities, which further shape (or filter) the spectrum of the modulated airflow signal. The resultant airflow is then radiated by the lips. Different types of sounds can be produced by manipulating the vibration pattern of the vocal folds, or the configuration of the vocal tract above the larynx by the articulations of the tongue, jaw, soft palate, and lips.

The *voice source* or *glottal source* is the volume velocity waveform that serves as the excitation of speech produced by the vocal folds. During voiced sounds, the oscillation of the vocal folds periodically interrupts the airow from the lungs and creates changes in air pressure (van den Berg, 1958; Kreiman and Sidtis, 2011). When no sound is being phonated, the vocal folds are usually open. To generate unvoiced sounds, the vocal folds are held apart, allowing the airow to pass through, and a constriction is formed at some point in the vocal tract to produce turbulence. To produce voiced sounds, the muscles which control the closing of the vocal folds (adductor muscles) bring the vocal folds together in order to provide resistance to the air pressure from the lungs. The air pressure below the closed vocal folds (subglottal pressure) forces the vocal folds to open, which allows airflow to pass through the glottis. Two factors contribute to the closing of the glottis again. The first is the elasticity of the tissue, which forces the vocal folds to regain their original configuration near the midline (the closed position). The second factor is the aerodynamic forces. One such force is described by the Bernoulli effect, which causes the drop of pressure between the vocal folds when airflow velocity increases. Another aerodynamic force occurs when vortices form in the airflow as it exits the glottis, creating an additional negative pressure between the vocal folds (McGowan, 1988). Once the vocal folds are closed, the air pressure below them builds up again, and the vocal folds are blown apart to start the process again. This cycle is repeated many times in each second, and the duration of the cycle is called the "fundamental period" (T_0). Its frequency is referred to as the "fundamental frequency" ($F_0 = 1/T_0$), or pitch frequency.

From a speech analysis point of view, glottal source processing is examined for both modal and non-modal phonation types (Gordon and Ladefoged, 2001). The modal voice represents a mostly neutral phonation type with little variation from period to period in successive glottal cycles. It also assumes that there is significant excitation around the glottal closure instant (GCI). Non-modal phonation types involve significant variation in glottal source characteristics. For analysis purposes, the different phonation types in normal speech are broadly categorized as shown in Figure 2 (Gordon and Ladefoged, 2001).

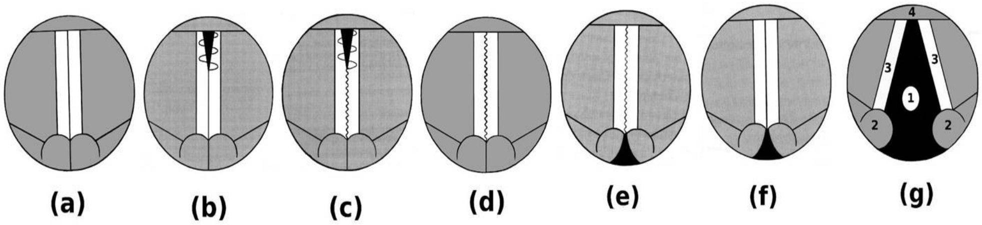

Figure 2: *Stylized glottal configurations for various phonation types: (a) Glottal Stop, (b) Creak, (c) Creaky voice, (d) Modal voice, (e) Breathy voice, (f) Whisper, (g) Voicelessness (1. Glottis, 2. Arytenoid cartilage, 3. Vocal fold and 4. Epiglottis)*

Note that only in the case of creaky voice, modal voice and breathy voice, is there vibration of the vocal folds, as shown by the corrugated line between the folds. Glottal source processing is typically used for these phonation types. Figure 2 clearly illustrates the other four types of phonation which have phonemic significance in the production of certain types of sound units in different languages.

High speed photography helps in understanding, for example, the differences in the vocal fold vibration between phonation types. Visual imaging of the vocal folds is, however, not possible during the natural production of continuous speech, when the vocal tract system also is time-varying. Moreover, visibly prominent features of glottal vibration may be less important from the point of view of speech production acoustics or speech perception. Electroglottography (EGG), in turn, which measures the time-varying impedance between electrodes placed on opposite sides of the neck and which can be measured (almost) non-invasively during production of continous speech, is unable to analyze the functioning of the airflow through the glottis. EGG is a measure of the vocal fold contact area, and may help as a reference (also called *ground truth*) when comparing, for example, pitch tracking or GCI detection methods.

Methods to derive physical or analytical models of the glottal source help us understand the significance of some features of the glottal vibration (Fant et al., 1985a). But these methods do not help in extracting actual features of glottal vibration from the speech signal, and hence do not help to identify the features that contribute to the specific voice qualities related to emotion or to some specific speech sounds, such as *trills*, for example.

According to the linear speech production model (Fant, 1970), speech signals are generated by filtering the voice source by the vocal tract transfer function. Many glottal source models have been proposed with varying levels of complexity, such as the Rosenberg (Rosenberg, 1971), Liljencrants-Fant (LF) (Fant et al., 1985b), Fujisaki-Ljungqvist (FL) (Fujisaki and Ljungqvist, 1986), and Rosenberg++ (R++) (Veldhuis, 1998) models. These models were derived from an analysis of physiological measurements. For illustration purposes, an example of one cycle of the glottal flow and its derivative in modal voice according to the widely-used LF model is given in Figure 3. The most important glottal phases are also indicated. The *open phase* refers to the timespan during which the vocal folds are opening, while the *return phase* is the period during which they are returning to their initial state. The open phase itself is divided into the *opening and closing phases* which are defined by the passage of the glottal flow by its maximum, as depicted in Figure 3. After the return phase, the vocal folds remain closed during the so-called *closed phase*.

The various aforementioned models of the glottal source describe a waveform similar to the one exhibited in Figure 3 using different analytical para-

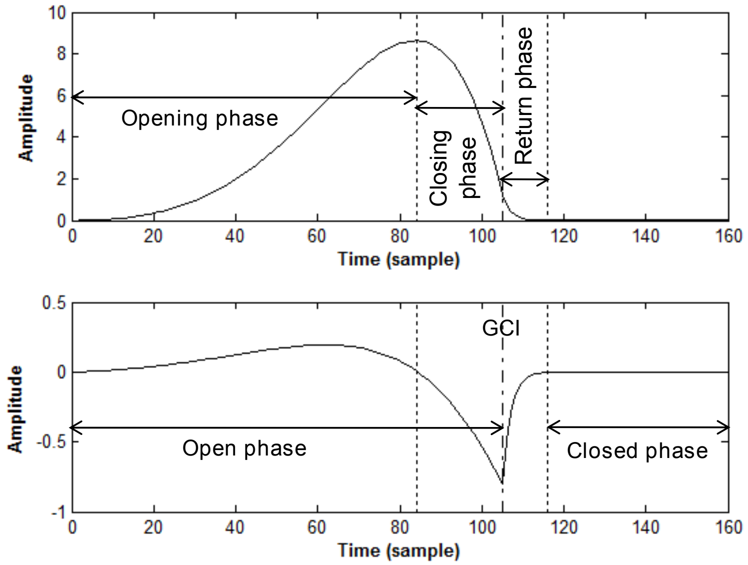

Figure 3: Typical waveforms, according to the Liljencrants-Fant (LF) model, of one cycle of: *(top)* the glottal flow, *(bottom)* the glottal flow derivative. The various phases of the glottal cycle, as well as the Glottal Closure Instant (GCI) are also indicated.

metric expressions. With three parameters, the Rosenberg trigonometric model has two separate functions for the opening and closing phases to represent the glottal flow volume velocity (Rosenberg, 1971). The LF and FL models represent the first derivative of the glottal volume velocity pulse, which incorporates lip radiation effects. The four-parameter LF model (Fant et al., 1985b) uses a combination of sinusoidal and exponential functions, and is commonly used in speech analysis and synthesis. With six parameters and polynomial functions, the FL model provides greater detail in modeling the glottal pulse shape, but the increased number of parameters also makes it more difficult to use in practice. The R++ model (Veldhuis, 1998) is computationally more efficient but perceptually equivalent when compared to the LF model. Motivated by high-speed images of the vocal folds, a four-parameter glottal flow model, denoted EE1, was introduced (Shue and Alwan, 2010). EE1 uses a combination of sinusoidal and exponential functions similar to the LF model, but with the ability to adjust the slopes of the opening and closing phases separately. Another glottal flow model, denoted EE2, by Chen et al. (2012) improves the EE1 model by redefining the model parameters (speed of opening and speed of closing) to allow for lower computational complexity, faster waveform generation, and more accurate pulse shape manipulation. The EE2 model was also used for automatic glottal flow estimation from acoustic speech signals. With the availability of more physiological data, we anticipate further improvements in source modeling.

3. Tools for Synchronization

Contrary to the conventional extraction of filter based features such as MFCC or LP parameters, which is carried out asynchronously, the estimation and parameterization of the glottal source (as will be explained in Section 4) sometimes involves the processing of speech frames whose duration is proportional to the pitch period, as well as knowledge of the GCI position. Note that the notion of synchrony must be understood here with regard to the glottal production (and not with absolute time): the analysis techniques are synchronous with the produced glottal cycles.

This section aims to provide a review of the existing tools necessary for proper synchronization: pitch tracking in Section 3.1, speech polarity detection (Section 3.2) and GCI determination in Section 3.3.

3.1. Pitch Tracking

The source-filter model of speech production assumes that speech signals can be modeled as an excitation signal filtered by a linear vocal-tract transfer function. The fundamental frequency (F_0) is defined as the inverse of the period of the excitation signal during the voicing state. Accurate F_0 tracking in quiet and in noise is important for several speech applications, such as speech analysis, coding and recognition.

Some F_0 tracking algorithms are based on the source-filter theory of speech production. They assume that F_0 is constant and the vocal tract transfer function is time invariant within a short period (10-20 milliseconds) of time. SIFT (Markel, 1972) applies inverse filtering to voiced speech to obtain the excitation signal, from which it estimates F_0 by computing the autocorrelation function. Cepstrum-based methods (e.g. (Noll, 1967)) separate the excitation from the vocal tract information in the cepstral domain by using a homomorphic transformation, the interval to the first dominant peak in the cepstrum is related to the fundamental period. RAPT (Talkin, 1995) and YAPPT (Kasi and Zahorian, 2002) generate F_0 candidates by extracting local maxima of the normalized cross correlation function, which is calculated over voiced speech. Praat (Boersma, 2001) calculates cross correlation or autocorrelation functions on the speech signal and regards local maxima as F_0 hypotheses. TEMPO (Kawahara et al., 1999) obtains F_0 candidates by evaluating a concept called fundamentalness of speech which achieves a maximum value when the AM and FM modulation magnitudes are minimized. YIN (de Cheveigne and Kawahara, 2002) uses the autocorrelation-based squared difference function and the cumulative mean normalized difference function calculated over voiced speech, with little post-processing, to acquire F_0 candidates. Yegnanarayana and Murty (2009a) obtain F_0 candidates by exploiting the impulse-like characteristics of excitation in glottal vibrations. Finally, Roux et al. (2007) simultaneously perform frame-wise F_0 candidate generation and time-direction smoothing.

In the multi-band method, a decision module is usually used to reconcile the F_0 candidates generated from different bands. Gold and Rabiner (1969) use measurements of peaks and valleys of voiced speech as input to six separate functions whose values are then processed by an F_0 estimator to obtain F_0 candidates. Lahat et al. (1987) calculate autocorrelation functions of the spectral magnitudes in different bands and then obtain F0 candidates by evaluating the local maxima of the functions. Sha et al. (2004) detect F0 candidates by minimizing the values of sinusoid-based error functions calculated on 4 frequency bands.

Multi-band methods (Lahat et al., 1987; Sha et al., 2004) typically retain F0 candidates obtained from the most reliable band, while those inspired by Licklider's theory of pitch perception use empirically-based 'soft-decisions' to merge the information from different bands (e.g. (de Cheveigne, 1991)). In (de Cheveigne, 1991), it is shown that integrating the Average Magnitude Difference Function (AMDF) values across different channels in the time domain can improve F0 estimation accuracy. They used correlograms to select reliable frequency bands, modeled F0 dynamics using a statistical approach, and then searched for the optimal F0 contour in an hidden Markov model(HMM) framework.

Some pitch tracking algorithms are designed to be noise robust. For example, a method based on the Summation of the Residual Harmonics (SRH) was proposed in (Drugman and Alwan, 2011). The SRH criterion exploits the harmonic structure of the Linear Prediction (LP) residual excitation both for pitch estimation, as well as for determining the voiced segments of speech. SRH was shown to be particularly robust to additive noise, leading to a significant improvement in adverse conditions over six representative state-of-the-art techniques. Finally, in (Chu and Alwan, 2012), a Statistical Algorithm for F_0 Estimation (SAFE) was proposed that utilizes a "soft-decision" method. A data-driven approach was used to learn how the noise affects the amplitude and location of the peaks in the Signal-to-Noise Ratio (SNR) spectra of clean voiced speech. The likelihoods of F_0 candidates were then obtained by evaluating the peaks in the SNR spectrum using the corresponding models learned from different bands.

3.2. Speech Polarity Detection

When a microphone is used to record speech, inverting its electrical connections will cause an inversion of the polarity of the acquired speech signals. The origin of a polarity in the speech signal stems from the asymmetric glottal waveform exciting the vocal tract resonances. During the production of voiced sounds, the glottal source exhibits periodic discontinuities at GCIs (see Figure 3). As described by the models of the glottal source introduced in Section 2, the speech polarity is defined as being positive if the glottal flow derivative exhibits a negative peak at the GCI. Otherwise it is said to be negative.

The human ear is mostly insensitive to a polarity change (Sakaguchi et al., 2000). Nonetheless the performance of several speech processing techniques can be severely deteriorated if the signal polarity is erroneous. This is the case for the majority of the GCI detection algorithms which will be described in Section 3.3, as well as for the methods of glottal source estimation and parameterization which will be explained in Section 4. Determining the speech polarity is then a necessary preliminary step to ensure that the aforementioned techniques work properly.

Several approaches have been designed for the automatic detection of the polarity of a speech signal. The idea of the Gradient of the Spurious Glottal Waveforms (GSGW, (Ding and Campbell, 1998)) technique is to investigate the discontinuity at the GCI in an estimated glottal waveform, whose sign is dependent upon the speech polarity. In the Phase Cut (PC, (Saratxaga et al., 2009)) method, the instant where the first two harmonics are in phase (and which should correspond roughly to the GCI) is first determined. If the phase value where they intersect is close to 0 (respectively π), this is expected to be due to a positive (respectively negative) peak in the excitation (Saratxaga et al., 2009). The Relative Phase Shift (RPS, (Saratxaga et al., 2009)) technique is derived from PC. It makes use of Relative Phase Shifts (RPS's), defined as the phase jumps between two consecutive harmonics.

When the excitation exhibits a positive peak, RPS's have a smooth structure across frequency. This smoothness is shown to be dramatically sensitive to a polarity inversion (Saratxaga et al., 2009). The Oscillating Moments-based Polarity Detection (OMPD) algorithm (Drugman and Dutoit, 2012a) first calculates statistical moments of the speech signal using a sample-by-sample sliding window. The resulting signals (called oscillating moments) have the property to oscillate at the local fundamental frequency when the window length is fixed properly. The key idea of OMPD is to rely on the observation that the phase shift between an even and odd-order oscillating moment is polarity sensitive. Finally, the Residual Excitation Skewness (RESKEW, (Drugman, 2013)) approach exploits the statistical skewness of two excitation signals: the LP residual, and a rough approximation of the glottal source. Indeed, since the skewness is known to be a measure of the asymmetry of a probability density function, it is used here as an estimator of the asymmetry of the glottal excitation. RESKEW has been shown to outperform all the techniques mentioned above in a test involving 10 large speech corpora, and to be superior in terms of computational load as well as robustness to noise and reverberation (Drugman, 2013).

3.3. Glottal Closure Instant Detection

The acoustic pressure variations caused by the glottal vibration to the airflow from the lungs can be viewed as the major excitation of the speech production system. This excitation signal is filtered by the response of the vocal tract system to generate the speech signal. If the excitation component is merely impulse-like, then the speech signal corresponds to the response of the vocal tract system. The nature of the impulse in terms of its sharpness is perceptually important and gives an indication of the strength or loudness (Seshadri and Yegnanarayana, 2009). In addition, the behavior of the glottal return phase, and the presence of other secondary excitations, also contribute to the voice quality of the sound.

It is interesting to note that knowledge of GCIs is useful in several speech analysis situations, such as detection of regions of glottal activity (Murty et al., 2009), estimation of F_0 (Yegnanarayana and Murty, 2009b), estimation of formant frequencies (Joseph et al., 2006), characterization of loudness of speech (Seshadri and Yegnanarayana, 2009), analysis of laugh signals (Kumar et al., 2009), and pitch extraction from multi-speaker data (Murty, 2009). Also, GCI-based analysis of speech can be used in several applications such as time delay estimation (Yegnanarayana et al., 2005), determination of number of speakers from mixed signals (Swamy et al., 2007), speech enhancement in single and multichannel cases (Prasanna, 2004), multi-speaker separation and prosody modification (Rao and Yegnanarayana, 2006), or speech synthesis (Drugman and Dutoit, 2012b).

Several reviews on GCI extraction can be found in the literature. Drugman gives a review focused on recent GCI detection methods (Drugman et al., 2012c). In another review, GCI detection over a range of voice qualities was discussed, with emphasis on GCI detection in modal and non-modal phonation (Kane and Gobl, 2013b). The non-modal phonation displays varying glottal source characteristics which include voices like creaky, breathy, tense, harsh and falsetto.

There are several phonation types involving glottal vibration that have distinct characteristics around the GCI, as shown in Figure 4. The figure shows the speech waveform, LP residual, EGG and difference EGG (dEGG) signals for different types of phonation, namely modal, breathy, tense, harsh, falsetto and creaky. The purpose of giving these cases is to show the differences in the features of glottal vibration, seen mostly in the dEGG. It is clear that the characteristics of the impulse-like excitation vary in each case, and that these characteristics are perceived by a human listener. But it is a challenge to extract such glottal source information from the speech signal.

The signal captured by the EGG allows the extraction of correct information about the GCI locations and the shape of the source signal around GCIs. However, it is preferable to have methods which extract GCIs directly from the speech signals, and use the GCI from EGG as reference for comparison. These methods may be grouped into the following five categories:

A: Methods which can exploit the property of impulse-like excitation at GCI.
B: Methods based on the properties of group-delay.
C: Methods based on predictability of all-pole linear predictor.
D: Methods based on the short-time energy of speech signal.
E: Methods based on a combination of several techniques.

Category A: The methods in this category pass the speech signal through a system which reduces the effects of the response of the vocal tract, at the same time preserving the impulse-like excitation property at GCIs. The zero frequency resonator (ZFR) is one such method where the speech signal is passed through a cascade of two ideal digital resonators located at 0Hz, so that the effects of all higher frequency resonances are reduced significantly

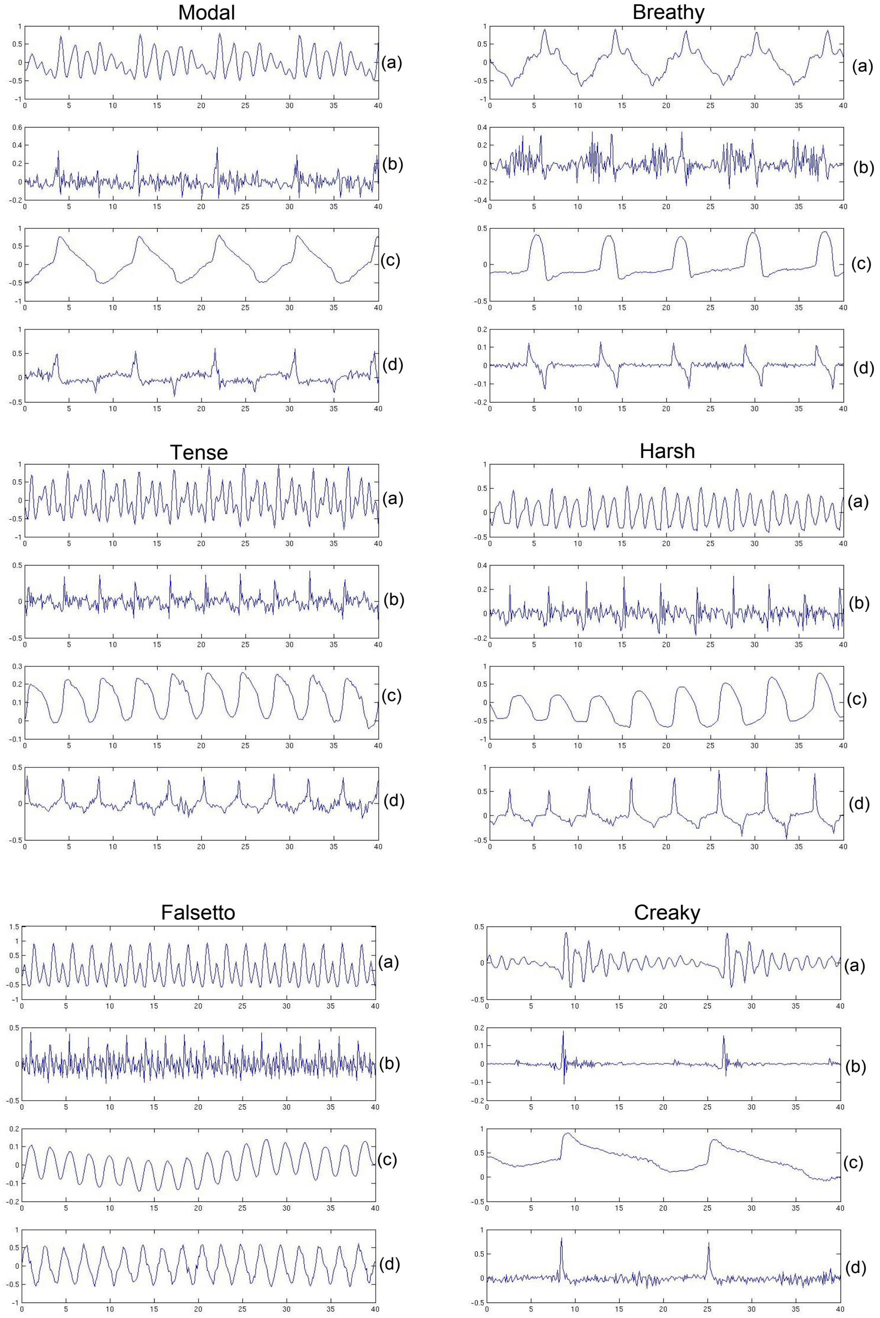

Figure 4: *Speech waveform (a), LP residual (b), EGG (c) and dEGG (d) signals for different phonation types: modal, breathy, tense, harsh, falsetto and creaky voice. The x-axis is the time (ms) and the y-axis is the amplitude.*

(Murty and Yegnanarayana, 2008). The trend in the ZFR output is removed by subtracting the average over a window of the size in the range of 1 to 2 pitch periods. The resulting mean subtracted signal, called zero frequency filtered (ZFF) signal gives the GCI locations precisely at the instants of negative to positive zero crossings. Exploiting the property of singularities detection by the wavelet transform, the concept of lines of maximum amplitude (LoMA) across all the scales in the wavelet transform domain is proposed for GCI detection (Tuan and d'Alessandro, 1999).

Category B: Group-delay is the negative derivative of the phase of the Fourier transform of a signal. The average of the group-delay function for each segment of windowed signal corresponds to the delay of the impulse in the segment from the center of the segment. Thus the average of the group-delay function computed at every sampling instant has a zero value at the GCI, i.e., when the GCI is located at the center of the analysis segment (Smits and Yegnanarayana, 1995). Several refinements on this approach lead to algorithms like DYPSA, which uses the zero crossings of the phase slope function derived from the energy weighted group-delay to obtain the candidates for GCI, and these candidate GCIs are refined by employing a dynamic programming algorithm (Naylor et al., 2007).

Category C: Many methods for GCI detection rely on the discontinuities in the output of a linear prediction model of speech production (Strube, 1974b). GCI detection from the LP residual is improved by computing its Hilbert envelope (HE), where unambiguous peaks are obtained around GCIs for clean signals, and these peaks match with the sharp discontinuities seen in the differenced EGG (Ananthapadmanabha and Yegnanarayana, 1979). There are several refinements suggested to improve the GCI detection from the LP residual (Kane and Gobl, 2013b).

One of the difficulties in using the prediction error for GCI detection is that the LP residual often contains effects due to resonances of the vocal tract system, as the inverse filter does not cancel out the resonances completely. Moreover, the inverse filter emphasizes the low signal to noise ratio (SNR) regions of the high frequency, and thus reduces the SNR in the LP residual. This reduction in SNR results in the loss of resolution of the impulse-like behavior around the GCI.

Category D: Some of the early methods for GCI detection are based on short time energy of the speech signal or from features in its time-frequency representation (Jankowski et al., 1995; Ma and Willems, 1994). Note that the time-frequency representation or energy computation require block processing, and hence the GCIs cannot be detected accurately.

Category E: Currently methods are being explored which combine several different methods used for GCI detection. The Yet Another GCI Algorithm (YAGA) is one such method which uses iterative adaptive inverse filtering, wavelet analysis, the group-delay function and M-best dynamic programming (Thomas et al., 2012). The method was also used for estimating the glottal opening instants.

The SEDREAMS algorithm is another method which uses a mean-based signal to determine short intervals where GCIs are expected to occur and then assigns to each interval a more accurate estimation of the GCI location by inspecting the LP residual (Drugman and Dutoit, 2009; Drugman et al., 2012c). The SEDREAMS algorithm has been further refined to handle GCIs from signals produced from different phonation types (Kane and Gobl, 2013b). The refinement involves applying a dynamic programming method to select the optimal path on estimated GCI locations based on both the strength of the LP residual peak and a transition cost (from one GCI location to the next). A further post-processing step is used to reduce the number of false peaks. The post-processing uses the output of the LP residual through a resonator located at a frequency corresponding to mean F_0.

This brief review shows that there is continuing effort on finding methods for GCI detection from speech signals which are not only accurate (with reference to the ground truth provided by dEGG), but also work for speech signals produced with different phonation types and for paralinguistic signals like laugh, cough, etc. The biggest challenge is to develop methods for GCI detection for speech signals collected in a practical environment, such as in a room with additive noise and moderate reverberation. A further challenge is to detect the other minor excitations within a glottal cycle (such as the glottal opening instant or secondary excitation peaks), as all the excitations (major and minor) together contribute to the perception of voice quality in speech. Understanding these glottal features may help in identifying the components of glottal source needed for synthesis of expressive or emotional speech.

4. Glottal Source Estimation and Parameterization

Glottal-based analysis of speech production consists typically of two stages. In the first stage, glottal flow waveforms are estimated using any of the techniques described in Section 4.1. In the second stage, the estimated waveforms are parameterized by expressing their most important properties in a compressed numerical form (Section 4.2). These techniques generally require synchronization information such as knowledge of F0 or GCI locations, whose extraction has been studied in Section 3.

4.1. Glottal Source Estimation

The article by Miller (1959) is regarded as the first publication in which Glottal Inverse Filtering (GIF) is used as a method to estimate the glottal source. This study was followed by several publications by Fant and his colleagues (Fant, 1961; Fant and Sonesson, 1962), as well as by, for example, Mathews et al. (1961). In these early studies, the inverse model of the vocal tract consisted of lumped analog elements which were adjusted based on visual observations of the filter output via an oscilloscope. Already in these early GIF studies the tuning of the anti-resonances was done searching for settings that yielded a maximally flat closed phase of the glottal pulse form (Lindqvist-Gauffin, 1964). This criterion has since been used in several GIF studies for determining optimal inverse filter settings.

To the best of our knowledge, the first study utilizing digital signal processing (DSP) in the estimation of the glottal source was published by Oppenheim and Schafer (1968). Their work actually addressed a more general tool of DSP, the so-called homomorphic analysis in which the convolved signal components are transformed into additive components using cepstral analysis. Separation of speech into the glottal source and vocal tract was studied as an example of cepstral analysis, hence presenting the first DSP-based experimental results of GIF. Digital cepstral analysis has been later used by (Drugman et al., 2009b) for instance as the means to separate the glottal source from the vocal tract. Another early study using digital GIF analysis was published by Nakatsui and Suzuki (1970). Their investigation takes advantage of digital filtering as a means to implement the anti-resonances of the vocal tract inverse model.

Rothenberg (1973) studied a GIF technique that is based on inverse filtering the volume velocity waveform recorded in the oral cavity rather than the acoustic speech pressure signal captured in the free field outside the mouth. He introduced a special pneumotachograph mask which is a transducer capable of measuring the volume velocity at the mouth. The recorded signal was then subjected to inverse filtering, where analog antiresonances were determined using spectrographic analysis. The use of the pneumotachograph mask benefits from providing an estimate for the absolute DC level of the glottal air flow, an issue that cannot be achieved with GIF techniques that use the free field microphone recording as the input. Even though the original work reported by Rothenberg (1973) involved the use of analog filters in cancelling the vocal tract resonances, the use of the pneumotachograph mask has later been combined with digital inverse filtering (Granqvist et al., 2003).

One of the most widely used GIF methods, closed phase (CP) covariance analysis was proposed by Strube (1974a). The method was further developed a few years later by Wong et al. (1979). The CP analysis uses linear prediction (LP) with the covariance criterion as a tool to compute digital vocal tract models. By positioning the analysis window of LP in the glottal closed phase (i.e., the timespan during which there is no contribution from the excitation, as shown in Figure 3 in Section 2), the CP analysis method is capable of separating the glottal source and the vocal tract. In comparison to analog techniques, the use of LP introduced a notable improvement in GIF because the vocal tract model adjusts automatically to the underlying speech signal. The CP analysis has been successfully used in many voice production studies (Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986). Other investigations, however, have demonstrated that the method is sensitive to the extraction of the closed phase region, and even small errors might result in severe distortion of the estimated glottal excitation (Riegelsberger and Krishnamurthy, 1993; Yegnanarayana and Veldhuis, 1998). This drawback can be partly alleviated by using an EGG instead of the acoustic speech signal in estimating the closed phase region (Krishnamurthy and Childers, 1986). In addition, the performance of CP has been reported to improve by using speech samples from consecutive cycles (Plumpe et al., 1999; Yegnanarayana and Veldhuis, 1998), adaptive high-pass filtering (Akande and Murphy, 2005), and constrained LP in modeling the vocal tract (Alku et al., 2009).

The idea of computing a parametric all-pole model of the vocal tract jointly with a source model was proposed by Milenkovic (1986) as a basis of his GIF technique. This approach enables utilizing speech samples over the entire fundamental period in the optimization of the vocal tract, an issue which is not fulfilled in the basic form of the CP analysis. The joint optimization of the vocal tract and glottal source has since been used by several authors. Some of these works have utilized the autoregressive model with an exogenous (ARX) input, where the input has been represented in a pre-defined parametric form (see Section 2) by using, for example, the RK or the LF model of the voice source (Isaksson and Millnert, 1989; Kasuya et al., 1999; Frohlich et al., 2001; Fu and Murphy, 2006; Berezina et al., 2010; Ghosh and Narayanan, 2011).

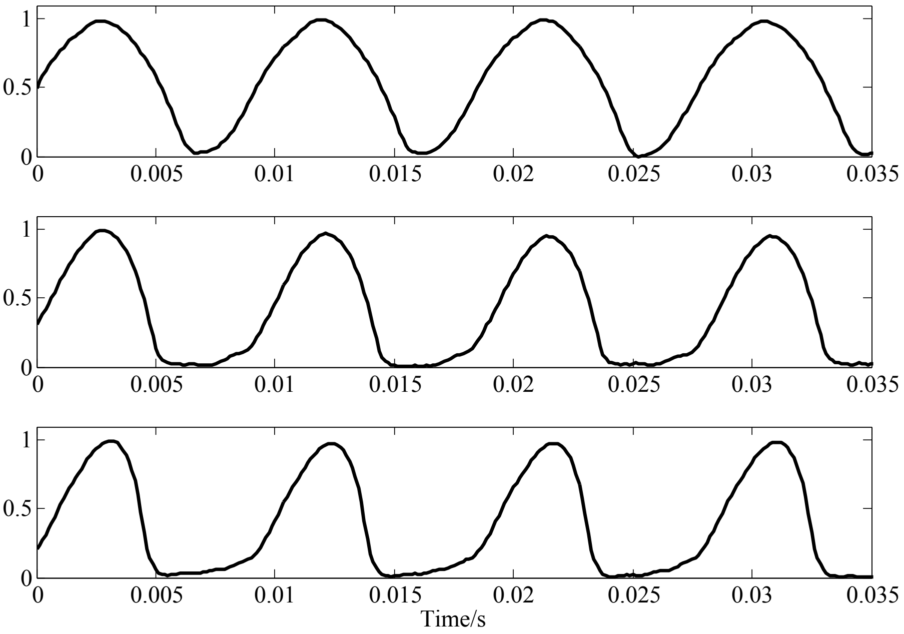

Figure 5: Examples of glottal flow estimates computed from natural speech using a GIF technique proposed in (Alku, 1992). Phonation type was varied from breathy (top panel) to modal (middle panel) and then to pressed (bottom panel). Changing of the glottal pulse from a soft and almost symmetric waveform into a skewed and asymmetric one can be seen when the phonation type is altered in the axis from breathy to pressed.

A straightforward and automatic GIF method, named Iterative Adaptive Inverse Filtering (IAIF), was proposed in (Alku, 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 modeling (DAP) that utilizes the Itakura-Saito distortion measure (El-Jaroudi and Makhoul, 1991). Examples of glottal flows estimated by IAIF are shown in Figure 5 for three different phonation types.

In parallel to the GIF techniques, non-parametric methods exploiting the maximum-phase properties of the speech signal have been proposed in (Bozkurt et al., 2005, 2007; Sturmel et al., 2007). 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 parts, corresponding to the glottal excitation and the vocal tract, based on their location with respect to the unit circle. A similar kind of separation based on causal-anticausal decomposition using the complex cepstrum has been proposed in (Drugman et al., 2011a, 2012a). An asynchronous scheme was proposed in (Drugman et al., 2009a). Separation based on phase characteristics (i.e. minimum-phase for the vocal tract and mixed-phase for the glottal source) was studied by Degottex et al. (2011a). Their method fits the phase spectrum of the LF model to that of the observed signal at harmonic components by minimizing the mean square phase difference. The technique proposed is different from the true GIF techniques (e.g. CP, IAIF) in the sense that the glottal source signal is not explicitly computed but model parameters are rather solved.

4.2. Glottal Source Parameterization

Once the glottal flow waveform has been estimated using any of the techniques presented in the previous section, it can be parameterized by expressing their most important properties in a compressed numerical form. In the following, some of the main parameterization techniques that have been developed are briefly described. We first discuss the time-domain methods and then the frequency-domain methods.

Parameterization of the time-domain glottal flow signals can be computed using time-based or amplitude-based methods. In the former, the most straightforward classical method is to compute time-duration ratios between different phases of the glottal flow pulse (Timcke et al., 1958; Monsen and Engebretson, 1977; Sundberg et al., 1999). Figure 6 illustrates an example of one cycle of the estimated glottal flow (top panel) and its derivative (bottom panel). The flow pulse is divided into three parts: the closed phase (T_c), the opening phase (T_o), and the closing phase (T_{cl}). The most widely used time-based parameters are defined as follows: open quotient ($OQ = \frac{T_o+T_{cl}}{T}$)), speed quotient ($SQ = \frac{T_o}{T_{cl}}$), and closing quotient ($ClQ = \frac{T_{cl}}{T}$), where the length of the fundamental period is denoted by $T = T_c + T_o + T_{cl}$.

Time-based measures described above are vulnerable to distortions that are present in glottal flow waveforms due to incomplete cancelling of formants by the inverse filter. Therefore, computation of the time-based parameters is sometimes performed by replacing the true opening and closure instants by the time instants when the glottal flow crosses a level which is set to a given value between the minimum and maximum amplitude of the glottal cycle (Dromey et al., 1992). In addition, time-based features of the glottal source can be quantified by measuring the amplitude-domain values of the glottal flow and its derivative (Fant, 1995; Alku and Vilkman, 1996; Alku et al., 2002). One such measure, the Normalized Amplitude Quotient (NAQ), was proposed by Alku et al. (2002). The computation of NAQ is done from two amplitude values (see Figure 6): the AC-amplitude of the flow (noted f_{AC}) and the negative peak amplitude d_{min} of the glottal flow derivative (see Figure 6). NAQ is then calculated as $NAQ = \frac{f_{AC}}{d_{min} \cdot T}$. Since these two amplitude measures are the largest values of the flow and its derivative in a glottal cycle, it is straightforward to extract even when the estimated glottal sources are distorted. In addition to NAQ, amplitude-based measures have been used in the parameterization of the voice source by Gobl and Chasaide (2003) who proposed methods to express three LF-based parameters (glottal frequency, skew parameter, and open quotient) in terms of amplitude-based measures.

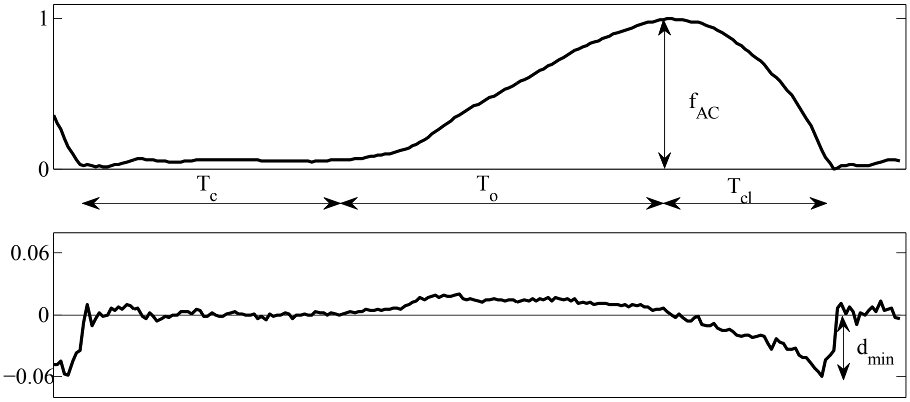

Figure 6: Time-based parameterization of the glottal source from the flow pulse (upper) and its derivative (lower panel). The flow pulse is divided into three parts: the closed phase (T_c), the opening phase (T_o), and the closing phase (T_{cl}).

When GIF is computed with the flow mask as proposed by Rothenberg (1973), it is possible to parameterize the amplitude-based properties of the time-domain waveforms of the glottal flow and its derivative. The most widely used amplitude parameters are the minimum flow (also called the DCoffset), 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).

It is also possible to parameterize the time-domain voice source by searching for a model waveform that matches the estimated glottal excitation (or its derivative) (Strik et al., 1993; Li et al., 2012). As exposed in Section 2, there are many such glottal source models (Fujisaki and Ljungqvist, 1986) that have been developed during the past three decades, ranging from the two-parameter polynomial model proposed by Klatt (1987) to more complex models such as the LF-model (Fant et al., 1985b). Time-domain reproduction of the glottal excitation estimated by inverse filtering is also possible by 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 (Drioli, 2005; Avanzini, 2008).

Instead of fitting the computed glottal flow with a pre-defined function, a parameterization scheme based on a data-driven approach was recently proposed by Gudnason et al. (2012). More specifically, a glottal flow estimate was first computed with IAIF and the obtained waveform was then processed with Principal Component Analysis (PCA) in order to achieve dimensionality reduction. The first few PCA components were then modelled with Gaussian Mixture Models (GMMs). The obtained GMMs were shown by Gudnason et al. (2012) to enable parameterizing source features, such as non-flatness of the closed phase, that traditional approaches fail to model. A new automatic parameterization method of the voice source was recently proposed by Kane and Gobl (2013a). The method is based on simulating the strategies that are used in manual optimization of the LF parameters. The technique proposed has three main parts. First, an exhaustive search is executed to get a group of most suitable parameter settings. A dynamic programming algorithm is then used to select the best path of the parameter values by considering two costs (target and transition) in the parameter trajectories of the modeled glottal pulses. As the final part, an optimization method is employed to refine the fit.

Frequency domain parameterization of the glottal source has been computed by measuring the spectral skewness with the so-called alpha ratio, which is the ratio between spectral energies below and above a certain frequency limit (typically 1.0 kHz) (Frokjaer-Jensen and Prytz, 1973). In addition, several studies have quantified the spectral slope of the glottal source by utilizing the level of the fundamental frequency and its harmonics. One such measure, called Harmonic Richness Factor (HRF), was developed by Childers and Lee (1991). HRF 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. In addition, levels of spectral harmonics have been used in parameterization of the glottal flow by Howell and Williams (1992), who measured the decay of the voice source spectrum by performing linear regression analysis over the first eight harmonics. Titze and Sundberg (1992) analyzed the spectral slope of the voice source of singers by computing the difference, denoted by H1-H2, between the amplitude of the fundamental and the second harmonic. A similar measure of the spectral tilt is computed directly from the radiated speech signal by correcting the effects of vocal tract filtering without computing the glottal flow by GIF (Iseli et al., 2007; Kreiman et al., 2012). Alku et al. (1997) proposed a frequency domain method based on fitting a low-order polynomial in the pitch synchronous source spectrum. Apart from measuring the spectral slope, it is also possible to quantify the glottal excitation by measuring the ratio between the harmonic and non-harmonic components (Murphy, 1999). This approach is justified especially in the analysis of 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.

5. Applicability of Glottal Source Processing

The analysis techniques discussed in Sections 3 and 4 allow the extraction of glottal source-related features. We will now see how this information can be incorporated in various voice technology applications: parametric speech synthesis in Section 5.1, expressive speech processing in Section 5.2, speaker recognition in Section 5.3, and biomedical applications in Section 5.4. The integration of glottal information in these applications is expected to be complementary to the vocal tract response.

5.1. Speech Synthesis

A great majority of existing techniques for parametric speech synthesis rely on a source-filter model. In this framework, two options are possible according to what is considered to be the source and the filter. In the first case, the source is the glottal flow as physiologically produced by the vocal folds, and the filter refers to the vocal tract response. Beyond the physiological motivation, this approach has the advantage of greater flexibility, as proper modifications of the glottal contribution are expected to reflect changes of voice quality. Nonetheless, the main drawback of this option is the requirement to reliably and accurately estimate and model the glottal source. In the second case, the filter corresponds to the spectral envelope of the speech signal and the excitation source is the residual signal obtained by inverse filtering after removing the spectral envelope contribution. The residual signal has the advantage to be easily obtained, however its amplitude spectrum is by definition flat and the information about the glottal spectral shaping is inextricably mixed in the filter component. Therefore its flexibility for speech modification is limited.

Using these two approaches, several methods have been proposed to improve the naturalness in HMM-based speech synthesis (Zen et al., 2009). Indeed, the basic vocoder used in HMM-based synthesis assumes the excitation signal to be a pulse train in voiced segments and white noise in unvoiced regions. This simple representation causes a typical *buzziness* in the generated speech, as was found in the old LP-based speech coders (Hedelin, 1986). To overcome this problem, a more elaborate source modeling is required. In (Yoshimura et al., 2001), a Mixed Excitation (ME) is proposed to model the residual signal. The ME is the sum of both periodic and aperiodic components which are controlled by bandpass voicing strengths. These latter parameters are fed to the HMMs during the training, and generated at synthesis time. In a similar way, a ME consisting of a set of high-order state-dependent filters derived through a closed-loop procedure was proposed in (Maia et al., 2007). In (Drugman et al., 2009e), a hybrid approach makes use of a codebook of pitch-synchronous residual frames which are selected at synthesis time, as in the Code Excited Linear Prediction (CELP, (Guerchi and Mermelstein, 2000)) method. In (Drugman et al., 2009d; Drugman and Dutoit, 2012b), the authors propose the Deterministic plus Stochastic Model (DSM) of the residual signal. The DSM consists of two contributions acting in two distinct spectral bands delimited by a maximum voiced frequency. Both components are extracted from an analysis performed on a speaker-dependent dataset of GCI-synchronous residual frames. The deterministic part models the low-frequency contents and arises from an orthonormal decomposition of these frames. As for the stochastic component, it is a high-frequency noise modulated both in time and frequency. The techniques modeling the residual signal have been shown to provide a significantly higher naturalness in

HMM-based speech synthesis, compared to the traditional pulse excitation.

In parallel, several other attempts have been made to integrate a glottal source modeling within HMM-based speech synthesis. The approach described in (Cabral et al., 2007) incorporates the LF model so as to reduce *buzziness* and enhance flexibility. A similar approach was proposed in (Lanchantin et al., 2010) where a new glottal source and vocal-tract separation method was used. Finally, a natural glottal pulse estimated by IAIF during a sustained vowel is used in the so-called GlottHMM approach presented in (Raitio et al., 2011). This glottal pulse is then further modified based on source spectral features and Harmonic-to-Noise Ratio (HNR) measures. Again, these latter methods were shown to outperform the traditional excitation in the HMM-based speech synthesis.

Besides the application of statistical parametric synthesis, several systems have targeted voice transformation by processing the excitation signal. Cabral et al. (2008), Degottex et al. (2011b) and Agiomyrgiannakis and Rosec (2009) proposed the use of the LF model to perform voice modifications (e.g. in terms of breathiness or tenseness of the generated speech). Several approaches have also focused on the manipulation of the excitation signal to carry out high-quality pitch modification (Cabral and Oliveira, 2005; Degottex et al., 2011b; Drugman and Dutoit, 2010a). Finally, the glottal source has also been employed in the context of voice conversion (i.e. with a specific target speaker in view) where, in addition to improving the segmental quality, it also offers the possibility to apply voice quality modifications (Childers, 1995; Pozo and Young, 2008).

5.2. Expressive Speech Processing

In expressive speech, the voice production significantly differs from the modal phonation. As articulation may completely be changed, the vocal tract function can be subject to important modifications. In parallel, much of the dynamic variation in voice quality is brought about by changes in phonation type, and hence changes in the glottal source signal (Laver, 1980). As a consequence, the production of expressive speech is also expected to be reflected by relevant alterations in the glottal contribution, as it was emphasized in Figure 4.

In (Monzo et al., 2007), the authors investigate the use of speech-related and glottal features to discriminate between five expressive speech styles: neutral, sad, happy, sensual and aggressive. The glottal parameters they investigate are the shimmer, jitter and the Glottal-to-Noise Excitation (GNE) ratio. These features are shown to provide interesting discrimination capabilities. In (Sun et al., 2009), the issue of differentiating emotions with a similar prosody is addressed by considering glottal parameters. Results show statistically significant differences in at least one glottal feature for all 30 emotion pairs where prosodic features did not show a significant difference. In (Tahon et al., 2012), the standard jitter and shimmer coefficients are complemented with the relaxation coefficient R_d and the Functions of Phase-Distortion (FPD). R_d is a parameter derived from the LF model, and it is known to quantify the tenseness in the voice. The FPD characterizes mainly the distortion of the glottal phase spectrum around its linear phase component. Results show that these glottal features are useful for the detection of the emotional valence (defined as the intrinsic attractiveness or aversiveness of an event, object, or situation). In (Szekely et al., 2011), the authors address the problem of expressive speech style clustering in order to develop high-quality text-to-speech synthesis from audiobooks. For this, they make use of a set of standard glottal features derived from the LF model (Oq, Sq and Rq) which are further clustered with a Self-Organising Feature Map (SOFM).

As a particular expressive style, the Lombard effect has received particular attention in the literature. The Lombard reflex refers to speech changes due to the immersion of the speaker in a noisy environment. In such a context, the modification of the residual signal is investigated in (Bapineedu et al., 2009) by considering features at the subsegmental level, namely, the strength of excitation and a loudness measure reflecting the sharpness of the impulse-like excitation at GCIs. In (Drugman and Dutoit, 2010b), the modifications of the glottal flow in Lombard speech are studied. For this, the glottal flow is estimated by CP analysis and parametrized by a set of time and spectral features. Significant and coherent changes of these parameters are observed as a function of the type and level of the surrounding noise. Another phonation type which involves a very specific glottal production has recently received particular attention: creaky voice (also referred to as vocal fry or laryngealisation). The glottal component during creaky voice is typically characterized by lower F0 values, a longer closed phase and secondary excitation peaks (Laver, 1980). The excitation-based automatic detection and synthesis of creaky voice have been respectively addressed in (Kane et al., 2003) and (Drugman et al., 2012b).

In addition to the aforementioned analysis and detection studies, several attempts have targeted the synthesis of expressive speech. The relevance of three speech components (spectral envelope, residual excitation and prosody) for synthesizing identifiable emotional speech has been estimated in (Barra et al., 2007). Results highlight the importance of transforming residual excitation for the identification of emotions that are not fully conveyed through prosodic means. In (Govind et al., 2011), the authors focus on the modification of the LP residual signal for emotion conversion. For this, the strength of excitation is modified by scaling the Hilbert envelope (HE) of the LP residual. The target emotion speech is finally synthesized using the modified excitation signal. The approach described in (Lorenzo-Trueba et al., 2012) aims to develop a HMM-based speech synthesis system with a controllable glottal source to manipulate the expressivity of the generated speech. The proposed approach relies on parameters of the GlottHMM vocoder proposed in (Raitio et al., 2011). As discussed in Section 5.1, this method estimates the glottal source using the IAIF technique and characterizes it by spectral parameters and HNR measures. Its viability is first analyzed by verifying that expressive nuances are captured by these features: recognition rates of 95% for styled speech and 82% for emotional speech are obtained. It is also shown that the method does not suffer from speaker and recording conditions bias. Finally, as a reminder, several techniques of parametric speech synthesis to modify the voice quality have been already presented at the end of Section 5.1.

5.3. Speaker Recognition

The baseline approach used in speaker recognition systems relies on standard MFCC-like features (Reynolds, 2002). As a consequence, most of the speaker-dependent information contained in the excitation source is discarded. However, as will be shown in this section, the use of the excitation information could be beneficial, as speakers with distinct larynx use their vocal folds in different ways. Furthermore, the speaker-dependent information contained in the glottal flow is expected to be complementary to the vocal tract information. Several methods have been proposed in the literature which differ on whether they use the residual signal or an estimate of the glottal source.

Several studies have underlined the fact that the residual signal conveys information regarding speaker identity. In (Yegnanarayana et al., 2001), the residual features are captured implicitly by a feedforward autoassociative neural network (AANN). This approach is further developed in (Prasanna et al., 2006) where it is shown that the residual information requires less training and testing data. The work presented in (Murty and Yegnanarayana, 2006) focuses on the use of the residual phase, defined as the phase of the analytic signal. The analytical signal is derived from the LP residual, making use of the Hilbert transform. In (Pati and Prasanna, 2008), the feature extraction from the residual signal is carried out by a non-parametric vector quantization. Chetouani et al. (2009) proposed the use of temporal and spectral features of the residual signal. Temporal characteristics are based on auto-regressive modeling, making use of second and third-order statistics to account for the non-linear nature of speech signals. As spectral features, the authors exploit a filter bank method measuring the spectral flatness over the sub-bands. All aforementioned approaches lead to the same conclusion: although the residual signal does not perform better than MFCC features, its complementary nature is demonstrated and its combination with MFCCs yields an improvement. Finally, the approach described in (Drugman and Dutoit, 2010c) and (Drugman and Dutoit, 2012b) proposes so-called *glottal signatures* derived from the Deterministic plus Stochastic Model (DSM) of the residual signal. Each speaker is characterized by two signatures corresponding to both the deterministic and stochastic components. This technique was shown to clearly outperform other speaker identification approaches based on glottal features.

Glottal flow estimates have also been proven to be useful for speaker recognition tasks. The first few attempts were made in (Plumpe et al., 1999). For this, the authors propose to estimate the glottal flow derivative using a variant of the CP analysis method (see Section 4.1). Glottal source estimates are modeled using the LF model to capture its coarse structure, while the fine structure of the flow derivative is represented through energy and perturbation measures. In (Slyh et al., 2004), the FL source model was used in conjunction with CP analysis to extract glottal features for a speaker recognition system. In (Gudnason and Brookes, 2008), a proper estimate of the vocal tract response is achieved by CP analysis. The glottal features then consist of the subtraction of the speech cepstral coefficients with those of the estimated vocal tract function. Finally, a similar approach relying on a cepstral representation of the voice source is proposed in (Kinnunen and Alku, 2009) and exploits an IAIF-based separation of the vocal tract and glottal components. Across all these studies, the use of glottal features alone did not provide better performance compared to the standard MFCCs, but their combination always led to a reduction of speaker recognition errors.

5.4. Biomedical Applications

At the crossroads between biomedical engineering and signal processing, glottal source processing has been demonstrated to be useful in several health-care applications. Its most straightforward application is probably for voice disorder detection. Indeed, speech pathologies are often associated with a dysfunction of the vocal folds, for example, oedema, or a polyp or nodule. Features describing the glottal behaviour are therefore of great interest for the automatic audio-based detection of voice pathologies. Jitter and shimmer are two typical measures employed to characterize irregularities in a quasi-periodic signal (Lieberman, 1963). Jitter refers to the change in the duration of consecutive glottal cycles, while shimmer reflects the changes in amplitude of consecutive glottal cycles. The study described in (Silva et al., 2009) focused on a comparative evaluation of different methods to estimate the amount of jitter in speech signals with a target on their ability to detect pathological voices. Results highlight significant differences in the performance of the various algorithms. In (Vasilakis and Stylianou, 2009), short-term jitter estimates are provided by the spectral jitter estimator (SJE), which is based on a mathematical description of the jitter phenomenon. SJE is proposed in order to discriminate voice pathologies in continuous speech. Jitter values were found to confirm studies showing a decrease of jitter with increasing fundamental frequencies, as well as a more frequent presence of high jitter values in the case of pathological voices. A variant of the traditional Harmonic-to-Noise Ratio (HNR) measure, called glottal-related HNR (GHNR') is proposed in (Murphy et al., 2008) to overcome the F0 dependency in the usual HNR formulation. For voice pathology discrimination, GHNR' is shown to provide statistically significant differentiating power over a conventional HNR estimator. The study described in (Drugman et al., 2009c) investigates the use of the glottal source estimation as a means to detect voice disorders. Three sets of features are proposed, depending on whether they are related to the speech or the glottal signal, or to prosody. Results indicate that speech and glottal-based features are relatively complementary, while they present some synergy with prosodic characteristics. The combination of glottal and speech-based features provides higher discrimination abilities. In (Gomez-Vilda et al., 2009), a biometric signature based on the power spectral density of the glottal source is presented. The detection capability is illustrated on a case study to determine a subject's voice condition in a pre- and post-surgical evaluation. Several phase-based features are proposed in (Drugman et al., 2011b) to detect voice disorders. The phase information is mostly linked with the glottal production, and it is shown to be particularly well suited to highlight irregularities of phonation compared to its magnitude counterpart.

It should also be noted that the use of glottal source processing in medical applications is not only restricted to pathological voices but the technologies developed can also applied to study such healthy voices which are in danger of becoming disordered due to vocal loading (Lauri et al., 1997; Vilkman et al., 1997). Vocal loading refers to prolonged use of voice in combination with additional environmental factors such as background noise and poor air quality (Vilkman, 2004). Vocal loading is likely to affect especially people, such as teachers, who use voice extensively in their daily work. Since the number of employees working in voice-intensive occupations is increasing, the occupational voice care will become an increasingly important application area for glottal source processing.

Glottal characterization has also been shown to be helpful in another biomedical problem: the classification of clinical depression in speech. In (Ozdas et al., 2004), the vocal jitter and glottal flow spectrum are proposed as possible cues for assessment of a patient's risk of committing suicide. Three groups of 10 subjects each are considered: high-risk near-term suicidal patients, major depressed patients, and non-depressed control subjects. The mean vocal jitter is found to be a significant discriminator only between suicidal and non-depressed control groups, while the slope of the glottal flow spectrum is a significant discriminator among all three groups. The approach presented in (Moore et al., 2008) also addresses the issue of discriminating depressed speech from normal speech. It is shown that the combination of glottal and prosodic features produces better discrimination than the combination of prosodic and vocal tract features. It is also underlined that glottal descriptors are vital components of vocal affect analysis. Toward the goal of objective monitoring of depression severity, (Quatieri and Malyska, 2012) investigates vocal-source biomarkers for depression. These biomarkers are specifically source features that may relate to precision in motor control, including vocal-fold shimmer and jitter, degree of aspiration and fundamental frequency dynamics. Results emphasize the statistical relationships of these vocal-source biomarkers with psychomotor activity, as well as with depression severity.

The use of the glottal source can also be of interest in applications for laryngectomees. Patients having undergone total laryngectomy cannot produce speech sounds in a conventional manner because their vocal folds have been removed. Gaining a new voice is then the major goal of the post surgery process. Several approaches have targeted the resynthesis of an enhanced version of alaryngeal speech, in order to improve its quality and intelligibility. For this purpose, it is required to re-create an artifical excitation signal based on our understanding of glottal production. This is generally achieved by using the LF model, as in the two following studies. In (Qi et al., 1995), the authors resynthesize female alaryngeal words with a synthetic glottal waveform and with smoothed and raised F0. It is shown that the replacement of the glottal waveform and F0 smoothing alone produces most significant enhancement, while increasing the average F0 leads to less dramatic improvement. The speech repair system proposed in (del Pozo and Young, 2006) resynthesizes alaryngeal speech using a synthetic glottal waveform, reduces its jitter and shimmer and applies a spectral smoothing and tilt correction algorithm. A subjective assessment reveals a reduction of the perceived breathiness and harshness of the voice. Finally, the solution described in (Sharifzadeh et al., 2010) interestingly focuses on the speech reconstruction from whispered voice, and proposes a modified version of the CELP vocoder.

Recent studies (such as (Tsanas et al., 2012)) have shown that dysphonia measures (similar to those used for voice pathology assessment) outperform state-of-the-art results for the detection of Parkinson's disease (PD). These features complement existing algorithms in maximizing the ability of the classifiers to discriminate healthy controls from PD subjects. This approach is a promising step toward the early non-invasive diagnostic of PD.

6. Conclusion

The goal of this paper was to provide a comprehensive review of the advances made in glottal source processing. Starting from the fundamental and necessary tools of synchronization (pitch tracking and GCI detection), we then presented methods for glottal flow estimation and parameterization. Finally, the last part of this paper looked at how the glottal-based information can be successfully integrated in some voice technology applications: speech synthesis, expressive analysis of speech, speaker recognition, and biomedical applications.

The advantage of glottal source processing is that the excitation signal is expected to be highly complementary to features conventionally used in current voice technology systems, and which mainly characterize the vocal tract response. However, this is done at the expense of an increase of the complexity of analysis techniques. Indeed, these latter techniques require the precise knowledge of the fundamental frequency and/or of the GCI positions, in contrast to conventional asynchronous (with regard to the glottal production) MFCC or LP-like feature extraction scheme. In addition, since neither the vocal tract response nor the glottal contribution are observable, estimation of the glottal source is a blind separation problem. In other words, although some techniques, such as EGG or high-speed digital imaging, can be used to gather information about the functioning of the vocal folds, the glottal flow produced by a natural speaker is unknown, and no reference is available for validating the accuracy of glottal source estimation methods. Nevertheless, EGG can be used as a ground truth for the development of pitch tracking and GCI detection algorithms.

The weakest point of current glottal source processing algorithms is related to their lack of robustness. Most of the approaches developed so far focus on the analysis of speech signals collected in controlled situations: synthetic signals, sustained vowels, studio recordings with high-quality condenser microphones. Although some progress has been made to improve robustness, further studies are definitely needed for the estimation of glottal information from continuous speech produced with different speaking styles in realistic environments (noisy and/or reverberant). Glottal analysis of such data is challenging because of the large variety of the underlying speech features: speech might be, for example, expressive, produced by shouting or by talkers, such as children using high pitch. To cope with this kind of extensive dynamics in acoustical speech features, glottal source processing must be made more robust to high F0, variations in aperiodicity, and should deal with the non-linear effects bewteen the vocal tract and the glottal flow. These research questions are important to be studied because success achieved can lead, for example, to improved speech quality and intelligibility when technologies developed are used in statistical parametric speech synthesis.

7. Acknowledgements

Thomas Drugman is supported by FNRS. The preparation of the article has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287678 (the Simple4all program) and from the Academy of Finland (project no. 135003 and 256961). Abeer Alwan is supported by the NSF.

References

Agiomyrgiannakis, Y., Rosec, O., 2009. ARX-LF-based source-filter methods for voice modification and transformation. ICASSP, 3589–3592.

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

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

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

Alku, P., Backstrom, T., Vilkman, E., 2002. Normalized amplitude quotient for parameterization of the glottal flow. Journal of the Acoustical Society of America 112, 701–710.

Alku, P., Magi, C., Yrttiaho, S., Backstrom, T., Story, B., 2009. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. Journal of the Acoustical Society of America 120, 3289–3305.

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

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

Ananthapadmanabha, T., Yegnanarayana, B., 1979. Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Speech Audio Processing 27, 309–319.

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

Bapineedu, G., Avinash, B., Gangashetty, S., Yegnanarayana, B., 2009. Analysis of Lombard speech using excitation source information. IEEE Trans. on Audio Speech and Language Processing, 1091–1094.

Barra, R., Montero, J. M., Macias-guarasa, J., Gutierrez-arriola, J., Ferreiros, J., Pardo, J. M., 2007. On the limitations of voice conversion techniques in emotion identification tasks. Interspeech.

Berezina, M., Rudoy, D., Wolfe, P., 2010. Autoregressive modeling of voiced speech. ICASSP, 5042–5045.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot International 5 (9), 341–345.

Bozkurt, B., Couvreur, L., Dutoit, T., 2007. Chirp group delay analysis of speech signals. Speech Communication 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 Sig.Pro. Letters 12, 344–347.

Cabral, J., Oliveira, L., 2005. Pitch-Synchronous Time-Scaling for Prosodic and Voice Quality Transformations. Interspeech, 1137–1140.

Cabral, J., Renals, S., Richmond, K., Yamagishi, J., 2007. Towards an improved modeling of the glottal source in statistical parametric speech synthesis. ISCA SSW6.

Cabral, J., Renals, S., Richmond, K., Yamagishi, J., 2008. Glottal spectral separation for parametric speech synthesis. Interspeech, 1829–1832.

Chen, G., Shue, Y.-L., Kreiman, J., Alwan, A., 2012. Estimating the voice source in noise. In: Interspeech.

Chetouani, M., Faundez-Zanuy, M., Gas, B., Zarader, J., 2009. Investigation on lp-residual representations for speaker identification. Pattern Recognition 42 (3), 487–494.

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

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

Chu, W., Alwan, A., 2012. Safe: A statistical approach to f0 estimation under clean and noisy conditions. IEEE Trans. on Audio Speech and Language Processing 20 (3), 933–944.

de Cheveigne, A., 1991. Speech f0 extraction based on lickliders pitch perception model. ICPhS, 218–221.

de Cheveigne, A., Kawahara, H., 2002. Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America 111 (4), 1917–1930.

Degottex, G., Roebel, A., Rodet, X., 2011a. Phase minimization for glottal model estimation. IEEE Trans. on Audio Speech and Language Processing 19, 1080–1090.

Degottex, G., Roebel, A., Rodet, X., 2011b. Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter. ICASSP, 5128–5131.

del Pozo, A., Young, S., 2006. Continuous tracheoesophageal speech repair. EUSIPCO.

Ding, W., Campbell, N., 1998. Determining polarity of speech signals based on gradient of spurious glottal waveforms. ICASSP, 857–860.

Drioli, C., 2005. A flow waveform-matched low-dimensional glottal model based on physical knowledge. Journal of the Acoustical Society of America 117, 3184–3195.

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

Drugman, T., 2013. Residual excitation skewness for automatic speech polarity detection. IEEE Sig. Pro. Letters 20 (4), 387–390.

Drugman, T., Alwan, A., 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. Interspeech, 1973–1976.

Drugman, T., Bozkurt, B., Dutoit, T., 2009a. Chirp decomposition of speech signals for glottal source estimation. ISCA Workshop on Non-linear Speech Processing.

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

Drugman, T., Bozkurt, B., Dutoit, T., 2011a. Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. Speech Communication 53, 855–866.

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

Drugman, T., Dubuisson, T., Dutoit, T., 2009c. On the mutual information between source and filter contributions for voice pathology detection. Interspeech, 1463–1466.

Drugman, T., Dubuisson, T., Dutoit, T., 2011b. Phase-based information for voice pathology detection. ICASSP, 4612–4615.

Drugman, T., Dutoit, T., 2009. Glottal closure and opening instant detection from speech signals. Interspeech, 2891–2894.

Drugman, T., Dutoit, T., 2010a. A comparative evaluation of pitch modification techniques. EUSIPCO.

Drugman, T., Dutoit, T., 2010b. Glottal-based analysis of the Lombard effect. Interspeech, 2610–2613.

Drugman, T., Dutoit, T., 2010c. On the potential of glottal signatures for speaker recognition. Interspeech.

Drugman, T., Dutoit, T., 2012a. Detecting speech polarity with high-order statistics. Cognitive Computation Journal.

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

Drugman, T., Kane, J., Gobl, C., 2012b. Modeling the creaky excitation for parametric speech synthesis. Interspeech.

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

Drugman, T., Wilfart, G., Dutoit, T., 2009d. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. Interspeech.

Drugman, T., Wilfart, G., Moinet, A., Dutoit, T., 2009e. Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. ICASSP, 3793–3796.

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

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

Fant, G., 1970. Acoustic theory of speech production, 2nd Edition. Mouton, The Hague, Paris, pp. 15-20.

Fant, G., 1995. The LF-model revisited. transformations and frequency domain analysis. Speech Transmission Laboratory Quarterly Progress and Status Report 36, 119–156.

Fant, G., Liljencrants, J., Lin, Q., 1985a. A four-parameter model of glottal flow. STL-QPSR 26(4), 1–13.

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

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, 1–3.

Frohlich, M., Michaelis, D., Strube, H., 2001. Sim simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. Journal of the Acoustical Society of America 110, 479–488.

Frokjaer-Jensen, B., Prytz, S., 1973. Registration of voice quality. Bruel&Kjaer Technical Review 3, 3–17.

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

Fujisaki, H., Ljungqvist, M., 1986. Proposal and evaluation of models for the glottal source waveform. ICASSP, 1605–1608.

Ghosh, P., Narayanan, S., 2011. Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter. Speech Communication 53, 98–109.

Gobl, C., Chasaide, A. N., 2003. Amplitude-based source parameters for measuring voice quality. ISCA VOQUAL, 151–156.

Gold, B., Rabiner, L., 1969. Parallel processing techniques for estimating pitch periods of speech in the time domain. Journal of the Acoustical Society of America 46 (2B), 442–448.

Gomez-Vilda, P., Fernandez-Baillo, R., et al., V. R.-B., 2009. Glottal source biometrical signature for voice pathology detection. Speech Communication 51 (9), 759–781.

Gordon, M., Ladefoged, P., 2001. Phonation types: a cross-linguistic overview. Journal of Phonetics 29 (4), 383–406.

Govind, D., Prasanna, S., Yegnanarayana, B., 2011. Neutral to target emotion conversion using source and suprasegmental information. Interspeech, 2969–2972.

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

Gudnason, J., Brookes, M., 2008. Voice source cepstrum coefficients for speaker identification. ICASSP, 4821–4824.

Gudnason, J., Thomas, M., Ellis, D., Naylor, P., 2012. Data-driven voice source waveform analysis and synthesis. Speech Communication 54, 199–211.

Guerchi, D., Mermelstein, P., 2000. Low-rate quantization of spectral information in a 4 kb/spitch-synchronous CELP coder. IEEE Workshop on speech coding, 111–113.

Hedelin, P., 1986. High quality glottal lpc-vocoding. ICASSP 11, 465–468.

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

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

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

Isshiki, N., 1981. Vocal efficiency index. K.N. Stevens and M. Hirano (Eds.), Vocal Fold Physiology, Tokyo, 193–203.

Jankowski, C., Quatieri, T., Reynolds, D., May 1995. Measuring fine structure in speech: Application to speaker identification. In: ICASSP. Detroit, MI, USA, pp. 325–328.

Joseph, M. A., Guruprasad, S., B, Y., Sep. 2006. Extracting formants from short segments using group delay functions. In: Interspeech. Pittsburgh, USA, pp. 1009–1012.

Kane, J., Drugman, T., Gobl, C., 2003. Improved automatic detection of creak. Computer Speech and Language 27 (4), 1028–1047.

Kane, J., Gobl, C., 2013a. Automatic manual user strategies for precise voice source analysis. Speech Communication 55, 397–414.

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

Kasi, K., Zahorian, S., 2002. Yet another algorithm for pitch tracking. ICASSP 1, 361–364.

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. Int. Congress of Phonetic Sciences, 2505–2512.

Kawahara, H., Katayose, H., de Chevigne, A., Patterson, R., 1999. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. Eurospeech 6, 2781–2784.

Kinnunen, T., Alku, P., 2009. On separating glottal source and vocal tract information in telephony speaker verification. ICASSP, 4545–4548.

Klatt, D., 1987. Review of text-to-speech conversion for english. Journal of the Acoustical Society of America 82, 737–793.

Kreiman, J., Shue, Y., Chen, G., Iseli, M., Gerratt, B., Neubauer, J., Alwan, A., 2012. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. Journal of the Acoustical Society of America 132, 2625–2632.

Kreiman, J., Sidtis, D. V. L., 2011. Foundations of voice studies: an interdisciplinary approach to voice production and perception. Wiley-Blackwell.

Krishnamurthy, A., Childers, D., 1986. Two-channel speech analysis. IEEE Trans. on Audio Speech and Signal Processing 34, 730–743.

Kumar, K. S., Reddy, M. S. H., Murty, K. S. R., Yegnanarayana, B., Sep. 2009. Analysis of laugh signals for detecting in continuous speech. In: Interspeech. pp. 1591–1594.

Lahat, M., Niederjohn, R., Krusback, D., 1987. A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. IEEE Trans. on Audio Speech and Signal Processing 35 (6), 741–750.

Lanchantin, P., Degottex, G., Rodet, X., 2010. A hmm-based speech synthesis system using a new glottal source and vocal-tract separation method. ICASSP, 4630–4633.

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

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

Li, H., Scaife, R., O'Brien, D., 2012. Automatic LF-model fitting to the glottal source waveform by extended kalman filtering. EUSIPCO, 2772–2776.

Lieberman, P., 1963. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. Journal of the Acoustical Society of America 35, 344–353.

Lindqvist-Gauffin, J., 1964. Inverse filtering. instrumentation and techniques. Speech Transmission Laboratory Quarterly Progress and Status Report 5, 1–4.

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

Ma, Y. K. C., Willems, L. F., April 1994. A Frobenius norm approach to glottal closure detection from the speech signal. IEEE Trans. Speech Audio Processing 2, 258–265.

Maia, R., Toda, T., Zen, H., Y.Nankaku, Tokuda, K., 2007. An excitation model for HMM-based speech synthesis based on residual modeling. ISCA SSW6.

Markel, J., 1972. The SIFT algorithm for fundamental frequency estimation. IEEE Trans. on Audio and Electroacoustics 20 (5), 367–377.

Mathews, M., Miller, J., David, E., 1961. Inverse filtering. instrumentation and techniques. Journal of the Acoustical Society of America 33, 179–186.

McGowan, R., 1988. An aeroacoustic approach to phonation. Journal of the Acoustical Society of America 83, 696–704.

Milenkovic, P., 1986. Glottal inverse filtering by joint estimation of an ar system with a linear input model. IEEE Trans. on Audio Speech and Signal Processing 34, 28–42.

Miller, R., 1959. Nature of the vocal cord wave. Journal of the Acoustical Society of America 31, 667–677.

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

Monzo, C., Alias, F., Iriondo, I., Gonzalovo, X., Planet, S., 2007. Discriminating expressive speech styles by voice quality parameterization. ICPhS, 2081–2084.

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.

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

Murphy, P., McGuigan, K., Walsh, M., Colreavy, M., 2008. Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals. Journal of the Acoustical Society of America 123 (3), 1642–1652.

Murty, K., Yegnanarayana, B., 2006. Combining evidence from residual phase and mfcc features for speaker recognition. IEEE Sig. Pro. Letters 13 (1), 52–55.

Murty, K. S. R., Mar. 2009. Significance of excitation source information for speech analysis. Phd thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras.

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

Murty, K. S. R., Yegnanarayana, B., Joseph, M. A., June 2009. Characterization of glottal activity from speech signals. IEEE Signal Process. Letters 16 (6), 469–472.

Nakatsui, M., Suzuki, J., 1970. Method of observation of glottal-source wave using digital inverse filtering in time domain. Journal of the Acoustical Society of America 47, 664–665.

Naylor, P. A., Kounoudes, A., Gudnason, J., Brookes, M., Jan. 2007. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans. Audio, Speech Lang. Process. 15 (1), 34–43.

Noll, A., 1967. Cepstrum pitch determination. Journal of the Acoustical Society of America 41 (2), 293–309.

Oppenheim, A., Schafer, R., 1968. Homomorphic analysis of speech. IEEE TAE 16, 221–226.

Ozdas, A., Shiavi, R., Silverman, S., Silverman, M., Wilkes, D., 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE Trans Biomed Eng. 51 (9), 1530–1540.

Pati, D., Prasanna, S., 2008. Non-parametric vector quantization of excitation source information for speaker recognition. TENCON, 1–4.

Plumpe, M., Quatieri, T., Reynolds, D., 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. on Audio Speech and Language Processing 7 (5), 569–586.

Pozo, A. D., Young, S., 2008. The linear transformation of lf glottal waveforms for voice conversion. Interspeech, 1457–1460.

Prasanna, S., Gupta, C., Yegnanarayana, B., 2006. Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Communication 48, 1243–1261.

Prasanna, S. R. M., Mar. 2004. Event based analysis of speech. Phd thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras.

Qi, Y., Weinberg, B., Bi, N., 1995. Enhancement of female esophageal and tracheoesophageal speech. Journal of the Acoustical Society of America 98, 2461–2465.

Quatieri, T., 2002. Discrete-time speech signal processing. Prentice-Hall.

Quatieri, T., Malyska, N., 2012. Vocal-source biomarkers for depression: A link to psychomotor activity. Interspeech.

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

Rao, K. S., Yegnanarayana, B., May 2006. Prosody modification using instants of significant excitation. IEEE Signal Process. Letters 14 (3), 972–980.

Reynolds, D., 2002. An overview of automatic speaker recognition technology. ICASSP 4, 4072–4075.

Riegelsberger, E., Krishnamurthy, A., 1993. Glottal source estimation: Methods of applying the LF-model to inverse filtering. ICASSP, 542–545.

Rosenberg, A., 1971. Effects of the glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Am. 49, 583–590.

Rothenberg, M., 1973. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. Journal of the Acoustical Society of America 53, 1632–1645.

Roux, J. L., Kameoka, H., Ono, N., de Cheveigne, A., , Sagayama, S., 2007. Single and multiple f0 contour estimation through parametric spectrogram modeling of speech in noisy environments. IEEE Trans. on Audio Speech and Language Processing 15 (4), 1135–1145.

Sakaguchi, S., Arai, T., Murahara, Y., 2000. The effect of polarity inversion of speech on human perception and data hiding as application. ICASSP 2, 917–920.

Saratxaga, I., Erro, D., Hernaez, I., Sainz, I., Navas, E., 2009. Use of harmonic phase information for polarity detection in speech signals. Interspeech, 1075–1078.

Seshadri, G., Yegnanarayana, B., Oct. 2009. Perceived loudness of speech based on the characteristics of excitation source. Journal of Acoustical Society of America 126 (4), 2061–2071.

Sha, F., Burgoyne, J., Saul, L., 2004. Multiband statistical learning for f0 estimation in speech. ICASSP 5, 661–664.

Sharifzadeh, H. R., McLoughlin, I., Ahmadi, F., 2010. Recontruction of normal sounding speech for laryngectomy patients through a modified celp codec. IEEE Trans. on Biomedical Engineering 57 (10).

Shue, Y.-L., Alwan, A., 2010. A new voice source model based on high-speed imaging and its application to voice source estimation. In: ICASSP. pp. 5134–5137.

Silva, D., Oliveira, L., Andrea, M., 2009. Jitter estimation algorithms for detection of pathological voices. EURASIP Journal on Advances in Signal Processing.

Slyh, R., Hansen, E., Anderson, T., 2004. Glottal modeling and closed-phase analysis for speaker recognition. ODYS, 315–322.

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

Strik, H., Cranen, B., Boves, L., 1993. Fitting a LF-model to inverse filtered signals. Eurospeech, 103–106.

Strube, H., 1974a. Determination of the instant of glottal closure from the speech wave. Journal of the Acoustical Society of America 56, 1625–1629.

Strube, H. W., 1974b. Determination of the instant of glottal closures from the speech wave. Journal of Acoustical Society of America 56, 1625–1629.

Sturmel, N., DAlessandro, C., Doval, B., 2007. A comparative evaluation of the zeros of z transform representation for voice source estimation. Interspeech, 558–561.

Sun, R., Moore, E., Torres, J., 2009. Investigating glottal parameters for differentiating emotional categories with similar prosodics. IEEE ICASSP, 4509–4512.

Sundberg, J., Andersson, M., Hultqvist, C., 1999. Effects of subglottal pressure on professional baritone singers' voice sources. Journal of the Acoustical Society of America 105, 1965–1971.

Swamy, R. K., Murty, K. S. R., Yegnanarayana, B., Jul. 2007. Determining number of speakers from multispeaker speech signals using excitation source information. IEEE Signal Process. Letters 14 (7), 481–484.

Szekely, E., Cabral, J., Cahill, P., Carson-Berndsen, J., 2011. Clustering expressive speech styles in audiobooks using glottal source parameters. Interspeech, 2409–2412.

Tahon, M., Degottex, G., Devillers, L., 2012. Usual voice quality features and glottal features for emotional valence detection. Speech Prosody.

Talkin, D., 1995. Robust algorithm for pitch tracking. Speech Coding and Synthesis, 497–518.

Thomas, M. R. P., Gudnason, J., Naylor, P. A., 2012. Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm. IEEE Trans. on Audio, Speech & Language Processing 20 (1), 82–91.

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

Titze, I., Sundberg, J., 1992. Vocal intensity in speakers and singers. Journal of the Acoustical Society of America 91, 2936–2946.

Tsanas, A., Little, M., Mcsharry, P., Spielman, J., Ramig, L., 2012. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. IEEE Trans. on Biomedical Engineering 59, 1264–1271.

Tuan, V. N., d'Alessandro, C., Sep. 1999. Robust glottal closure detection using the wavelet transform. In: European Conf. Speech Processing, Technology. Budapest, pp. 2805–2808.

van den Berg, J., 1958. Myoelastic-aerodynamic theory of voice production. J. Speech Hear. Res. 1, 227–244.

Vasilakis, M., Stylianou, Y., 2009. Voice pathology detection based on short-term jitter estimations in running speech. Folia Phoniatr Logop. 61 (3), 153–170.

Veeneman, D., BeMent, S., 1985. Automatic glottal inverse filtering from speech and electroglottographic signals. IEEE Trans. on Audio Speech and Signal Processing 33, 369–377.

Veldhuis, R., 1998. A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation. Journal of the Acoustical Society of America 103, 566–571.

Vilkman, E., 2004. Occupational safety and health aspects of voice and speech professions. Folia Phoniatrica et Logopaedica 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 Phoniatrica et Logopaedica 49, 247–263.

Walker, J., Murphy, P., 2007. A review of glottal waveform analysis. Springer Lecture Notes in Computer Science (LNCS) 4391, 1–21.

Wong, D., Markel, J., Gray, A., 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans. on Audio Speech and Signal Processing 27, 350–355.

Yegnanarayana, B., Murty, K., 2009a. Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. on Audio Speech and Language Processing 17 (4), 614–624.

Yegnanarayana, B., Murty, K. S. R., May 2009b. Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. on Audio, Speech, and Language Processing 17 (4), 614–624.

Yegnanarayana, B., Prasanna, S. R. M., Duraiswamy, R., Zotkin, D., Nov. 2005. Processing of reverberant speech for time-delay estimation. IEEE Trans. on Speech and Audio Processing 13 (6), 1110–1118.

Yegnanarayana, B., Reddy, K. S., Kishore, S., 2001. Source and system features for speaker recognition using aann models. ICASSP 1, 409–412.

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

Yoshimura, T., Tokuda, K., Masuko, T., Kitamura, T., 2001. Mixed-excitation for HMM-based speech synthesis. Eurospeech, 2259–2262.

Zen, H., Tokuda, K., Black, A., 2009. Statistical parametric speech synthesis. Speech Communication 51 (2), 1039–1064.