Computer Science Department
University of Crete

Institute of Computer Science
FO.R.T.H.

# Determining Glottal Closure and Opening Instants in Speech

Master's Thesis

**Christina Alexandra Lionoudaki**

Heraklion
March 2010

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF CRETE

**Determining Glottal Closure and Opening Instants in Speech**

Submitted to the Department of Computer Science in partial fulfillment of the requirements for the degree of Master of Science

March 26, 2010



Author

Lionoudaki Christina-Alexandra

Board of enquiry:
Supervisor

Stylianou Yannis
Associate Professor

Member

Tziritas Georgios
Professor

Member

Mouchtaris Athanasios
Assistant Professor

Accepted by
the Chairman of the
Graduate Studies Committee:

Panos Trahanias
Professor

Heraklion, March 2010

# Determining Glottal Closure and Opening Instants in Speech

**Christina Alexandra Lionoudaki**

Master's Thesis

Computer Science Department
University of Crete

## Abstract

Voice quality is a complex attribute of voice but one important aspect arises from the regularity and duration of the closed phase from vocal fold cycle to cycle. The determination of closed phase requires the accurate detection of glottal closure (GCI) and glottal opening (GOI) instant. In literature, many methods have been suggested on this direction employing either the Electroglottographic (EGG) or the speech signal. This work presents a robust algorithm for the detection of glottal instants from the EGG signal and a study on the interaction between Amplitude–Frequency components of speech and glottal phases.

The determination of GCIs and GOIs, is quite straightforward using Electroglottographic (EEG) signals. The derivative of EGG offers a simple way in detecting the important instances during the production of speech; the glottal closing and opening instants. In this thesis we suggest an alternative method to the simple derivative which is based on the spectral methods. Spectral methods provide an elegant way to conduct first and higher order derivatives on discrete time data, with high accuracy.

Furthermore, we introduce a new way to differentiate the EGG signal for estimating the main glottal instants. The gradient of electroglottographic signal is performed with a method referred to as "Slope Filtering". This approach shows to be robust in revealing the major peaks in the slope filtered EGG signal, even in cases where the quality of the EGG recordings is not good. Contrary to the simple derivative of the EGG signal, the peaks can be well distinguished and uniquely specified in the slope filtered signal. The proposed method exhibits high accuracy of voiced segments, including the onset and offset regions.

The derivation of glottal phases from speech signal has drawn great attention in recent years. A novel approach, relying on the speech signal, is proposed based on a Quasi–Harmonic (QHM) representation of speech. The adaptive QHM algorithm estimates the instantaneous AM–FM components of the speech signal. The extracted components, which are used for the reconstruction of the signal, are correlated to the glottal phases generated from the EGG signal. The AM component follows a steady pattern for each glottal phase, whereas the FM component shows low variations for various speakers.

Supervisor: Yannis Stylianou
Associate Professor
University of Crete

# Προσδιορισμός στιγμών κλεισίματος και ανοίγματος της γλωττίδας στη φωνή

**Χριστίνα Αλεξάνδρα Λιονουδάκη**

Μεταπτυχιακή Εργασία

Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

## Περίληψη

Η ποιότητα της φωνής είναι ένα σύνθετο χαρακτηριστικό, αλλά ένας σημαντικός παράγοντας προκύπτει από την ομαλότητα και τη διάρκεια της φάσης που οι φωνητικές χορδές παραμένουν κλειστές από κύκλο σε κύκλο. Ο προσδιορισμός του κλειστού αυτού διαστήματος απαιτεί την ακριβή ανίχνευση της στιγμής κλεισίματος (GCI) και της στιγμής ανοίγματος (GOI) της γλωττίδας. Στη βιβλιογραφία, πολλές μέθοδοι έχουν διεξαχθεί προς αυτή την κατεύθυνση που χρησιμοποιούν είτε το Ηλεκτρογλωττογραφικό (EGG) σήμα ή το σήμα φωνής. Η εργασία αυτή παρουσιάζει ένα αξιόπιστο αλγόριθμο για την ανίχνευση των γλωττιδικών στιγμών από το EGG σήμα και μια μελέτη σχετικά με την αλληλεπίδραση μεταξύ της Διαμόρφωσης του Πλάτους και της Συχνότητας του σήματος φωνής με τις γλωττιδικές φάσεις.

Ο προσδιορισμός των γλωττιδικών στιγμών είναι πολύ άμεσος χρησιμοποιώντας τα Ηλεκτρογλωττογραφικά σήματα. Η παράγωγος του EGG στο χρόνο προσφέρει έναν απλό τρόπο για τον εντοπισμό των σημαντικών στιγμών κατά τη διάρκεια της παραγωγής του λόγου: των στιγμών κλεισίματος και ανοίγματος της γλωττίδας. Στην παρούσα εργασία προτείνουμε μια εναλλακτική μέθοδο από την απλή παράγωγο, η οποία βασίζεται στις φασματικές μεθόδους. Οι φασματικοί μέθοδοι παρέχουν έναν καλαίσθητο τρόπο για τον υπολογισμό της πρώτης και μεγαλύτερων τάξεων παραγώγων για δεδομένα σχετικά με διακριτό χρόνο, με υψηλή ακρίβεια.

Επιπλέον, συστήνουμε ένα νέο τρόπο για την διαφόριση του EGG σήματος ως προς την εκτίμηση των κύριων γλωττιδικών στιγμών. Η κλίση του EGG σήματος προσεγγίζεται με μια μέθοδο που αναφέρεται ως "Slope Filtering". Αυτή η προσέγγιση δείχνει να είναι αξιόπιστη στην αποκάλυψη των κυρίων κορυφών του 'slope' φιλτραρισμένου EGG σήματος, ακόμη και σε περιπτώσεις όπου η ποιότητα των EGG καταγραφών δεν είναι καλή. Σε αντίθεση με την απλή παράγωγο του EGG σήματος, οι κορυφές μπορεί να διακρίνονται ξεκάθαρα και να ορίζονται μοναδικά στο 'slope' φιλτραρισμένο σήμα. Η προτεινόμενη μέθοδος παρουσιάζει υψηλή ακρίβεια στα έμφωνα, συμπεριλαμβανομένων και των περιοχών της έμβασης (onset) και έκβασης (offset) των περιοχών αυτών.

Η παραγωγή γλωττιδικών φάσεων από το σήμα ομιλίας έχει ιδιαίτερο ενδιαφέρον, μιας και πρόκειται για μη επεμβατική μέθοδο. Μια νέα προσέγγιση προτείνεται, η οποία στηρίζεται στο σήμα ομιλίας και βασίζεται στο περίπου Αρμονικό μοντέλο (QHM). Ο προσαρμοστικός QHM αλγόριθμος παρέχει εκτιμήσεις των στιγμιαίων AM–FM συνιστω-

σών του σήματος ομιλίας. Τα στοιχεία που εξάγονται, τα οποία χρησιμοποιούνται για την ανακατασκευή του σήματος, συγκρίνονται με τις γλωττιδικές φάσεις που παράγονται από το EGG σήμα. Η διαμόρφωση κατά πλάτος (AM) ακολουθεί ένα σταθερό πρότυπο για κάθε γλωττιδική φάση, ενώ η διαμόρφωση κατά συχνότητα (FM) δείχνει μικρές διαφοροποιήσεις από ομιλητή σε ομιλητή.

Επόπτης: Ιωάννης Στυλιανού
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Κρήτης

*Στους γονείς μου, Παύλο και Ελένη,*
*στην αδερφή μου, Εμμανουέλα,*
*και στον Χρήστο*

## Ευχαριστίες

*Καταρχήν θα ήθελα να ευχαριστήσω τον επόπτη της μεταπτυχιακής μου εργασίας Αναπληρωτή Καθηγητή κ.Ιωάννη Στυλιανού για την συνεχή καθοδήγηση και υποστήριξή του στο πλαίσιο της συνεργάσίας μας κατά την διάρκεια των μεταπτυχιακών μου σπουδών.*

*Στη συνέχεια θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή κ.Γεώργιο Τζιρίτα και τον Επίκουρο Καθηγητή κ.Αθανάσιο Μουχτάρη που συμπλήρωσαν ως μέλη την εξεταστική επιτροπή της μεταπτυχιακής εργασίας.*

*Επιπλέον θα ήθελα να ευχαριστήσω το Εργαστήριο Τηλεπικοινωνιών και Δικτύων του Ινστιτούτου Πληροφορικής του Ιδρύματος Τεχνολογίας και Έρευνας, το Εργαστήριο Ελέγχου Ποιότητας Φωνής και το Εργαστήριο Πολυμέσων του Πανεπιστημίου Κρήτης για την οικονομική υποστήριξη και τον διαθέσιμο εξοπλισμό.*

*Πολλές ευχαριστίες οφείλω στον υποψήφιο διδάκτορα Ιωάννη Πανταζή για την πολύτιμη καθοδήγηση προς την ολοκλήρωση της εργασίας. Επίσης ευχαριστώ για την ανεκτικότητα και συμπαράσταση τα πρωήν και νυν μέλη του εργαστήριου: Μίλτο Βασιλάκη, Ιωάννη Σφακιανάκη, Γιώργο Γκρέκα, Γιώργο Τζεδάκη, Μαρία Κουτσογιαννάκη, Μαίρη Αστρινάκη, Χρήστο Τζαγκαράκη.*

*Στη διάρκεια της μεταπτυχιακής μου εργασίας έλαβα αμέριστη υποστήριξη από τους συμφοιτητές μου: Γιώργο Μπαργιάννη, Κατερίνα Μπούτσικα, Ζωή Σεμπέπου, Γιώργο Τεσσέρη, Πάνο Τουρλάκη.*

*Επιπρόσθετα, ευχαριστώ τους μακροχρόνιους φίλους μου: Ελίζα Διαμαντίδη, Έλλη Βατσάκη, Μαίρη Μεντόνη, Αντώνη Μοτάκη, για την κατανόηση, συμπαράσταση που έδειξαν και στάθηκαν δίπλα μου όλο αυτό το διάστημα.*

*Τέλος θέλω να ευχαριστήσω θερμά τους γονείς μου, Παύλο και Ελένη, την αδερφή μου, Εμμανουέλα, καθώς και τον συντροφό μου, Χρήστο Τρουλάκη, για την ανιδιοτελή στήριξη και αγάπη τους σε όλα τα βήματα μου.*

*Σας ευχαριστώ πολύ!!!*

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

The voice quality depicts phonatory and resonatory characteristics such as harshness, breathiness, nasality. These features provide a sense of the perception of voice and of any existing pathology. Laryngeal differences are related to voice quality and consequently any effect in larynx, glottal airflow or vocal folds provides an additional information. Voice quality is a complex attribute of voice but one important aspect comes from the regularity and duration of the closed phase from vocal fold cycle to cycle.

Glottal Closure Instants correspond to the instants of significant excitation of the vocal tract. These particular time events reflect the moments of high energy in the glottal signal during voiced speech. The moment that vocal folds begin to close is called Glottal Closure Instant (will be referred as GCI) and the moment that vocal folds begin to open is called Glottal Opening Instant (will be referred as GOI).

The identification of the mentioned Instants and the determination of closing and opening phases of vocal folds offer a more precise picture of vocal folds' movement. The approximation of closing and opening phase and the relationship in duration of the closing and opening stages of the vocal fold vibratory cycle constitute descriptors and measures for qualitative (perceptual) and quantitative (physical) speech features. These measures contribute significantly in visual feedback purposes and evidence of improvement in vocal fold function after surgical intervention in existent pathology of the vocal folds. For instance, pathology such as organic lessions of vocal folds [10] and puberphonia in adults [11], where the variation in vocal fold contact is clearly observed between the closed and open phases of the vibratory cycle. Many pathological cases are associated with larynx where diagnosis cannot be executed without an invasive method, such as stroboscopy, photoglottography. Laryngeal pathology includes polypoid of the vocal folds, vocal fold pseudocyst, vocal fold hematoma, sarcoidosis, papilloma, unilateral vocal fold paralysis and others. It is of great importance to assess the intelligibility of a patient's speech before, during and at the end of therapy (with surgical intervention or not). The final display of the characteristics of vibratory cycle is of particular

value, not only in demonstrating change to the therapist and patient, but also in establishing efficacy of treatment. By inference, the determination of GCIs, GOIs and the duration of each phase may aid in diagnosis and assessment of treatment of laryngeal pathology.

A method for detecting glottal Instants is through the examination of the Electroglottographic (EGG) or Laryngographic signal. This is a direct method based on Lx/EGG waveform. Although, much work has been contributed in this area, more attention is given to the identification of GCIs, while the opening Instants are related to GCIs and not appreciated separately. The rapid rise in the waveform to a maximum on vocal fold approximation (the 'closing phase'), the gradual fall (the 'opening phase') and their duration is important and consequently the need for more accurate determination of GOIs is essential.

The derivative of EGG is widely applied with various approaches in this effort. The positive peak of the derivative corresponds to the closure Instant of EGG and the negative peak to the opening Instant. However, in several instances the peaks of the derivative are not easily detected, especially the negative ones. In these cases the derivative forms many negative peaks and the discrimination is complicated. Thus, the GCIs are firstly determined and then the GOIs are located to a certain distance from the first. Furthermore, many approaches impose a threshold on the differentiated EGG signal, which provide accurate results during voiced speech but are prone to errors at the onset and end of voicing. In some cases, peaks are doubled or imprecise, which points to special but not uncommon glottal configurations.

In addition, several techniques that estimate GCIs and GOIs extract information exclusively from speech signal. This is a field of growing interest, as an aspect of voice quality may be determined merely from speech signal. There are several areas of speech processing in which it is beneficial to be able to identify the GCIs and the closing phase intervals. Many speech processing algorithms are based in the detection of GCIs in voiced speech. Glottal-synchronous processing in speech synthesis, speaker characterisation for synthesis, low bit-rate speech coding and transmission, voice quality enhancement, voice transformation, prosodic speech modification, speech dereverberation, speech recognition, speech analysis, speaker normalization and recognition and fields of speech pathology and therapy apply detection of GCIs. Accurate identification of the closed phases allows the blind deconvolution of the vocal tract and glottal source through the use of closed phase analysis and modeling [46, 2, 7], without the use of additional equipment (Laryngograph and electrodes).

The last few years considerable methods have been proposed in estimating GCIs and GOIs from speech features. However, the presence of reverberation, noise and filtering by the vocal tract render Instants' estimation from real speech signal rather difficult to achieve compared with the results detected from EGG, so EGG-extracted features have been often referred to evaluate Instants' approximation from speech signal.

## 1.2 Contributions

The derivative of EGG offers a simple way to detect the important instances during the production of speech; the glottal closing and opening instants. In this work we suggest an alternative method to the simple derivative, that is usually used for generating the derivative of EGG signal, which is based on the spectral methods [53]. Spectral methods provide an elegant way to conduct first and higher order derivatives on discrete time data, with high accuracy. Experiments have shown that spectral methods provide slightly better results compared to the simple derivative approach, in terms of visibility of the major positive and negative peaks used for the detection of GCIs and GOIs.

In this thesis, we present a new way to differentiate the EGG signal for estimating the main glottal instants. We underline the insufficiency of the existing algorithms to estimate glottal opening instants. Each opening instant is principally associated to its distance from GCI. Furthermore, the gradient of electroglottographic signal is performed with a method referred to as "Slope Filtering", which was proposed in [55]. The EGG signal is filtered in a frame-by-frame basis by an FIR system consisting of a short impulse response, taking into account the vicinity of the samples around the center of each frame. This approach shows to be robust in revealing the major peaks in the slope filtered EGG signal, even in cases where the quality of the EGG recordings is not good. Then, the main glottal instants are easily detected using a simple thresholding approach. Contrary to the simple derivative of the EGG signal, the peaks can be well distinguished and uniquely specified in the slope filtered signal. The proposed method exhibits high accuracy of voiced segments, including the onset and offset regions.

Moreover, we apply an AM–FM decomposition method in order to reveal the relationship between the AM–FM components of speech signal and the glottal phases. The iterative AM–FM decomposition algorithm is based on the adaptive Quasi–Harmonic model (aQHM) [40]. The aQHM suggests a non-parametric AM-FM decomposition algorithm, which proceeds by successive adaptations of the decomposition basis functions to the characteristics of the underlying sine waves of the input signal. The extracted components, which are used for the reconstruction of the signal, are compared to the glottal phases generated from the EGG signal. The AM modulation follows a steady pattern for each glottal phase, whereas the FM component shows low variations for various speakers. The EGG signal is time–aligned with the speech signal and glottal Instants are derived from EGG signal. The glottal phases are depicted with the AM–FM components to derive the desired information of their variation within a glottal cycle.

## 1.3 Thesis Outline

The remaining chapters are organized as follows. The second chapter covers the background theory, in which the physiology of larynx and the interaction between the glottis and the vocal tract are introduced. In addition, the relation between the electrolaryngographic signal and the

glottal airflow is analyzed. The phases and Instants of EGG signal are thoroughly examined. The third chapter outlines the research that has been reported so far in the detection of glottal Closures and Openings from EGG signal or its derivative and from speech signal.

The next chapter refers to the spectral derivative of EGG signal and the "Slope Filtering" method. The applied methodology for the determination of glottal Instants from the EGG signal is described. The recorded database and the implemented techniques are analyzed. The results of the experiments are explicitly presented and conclusions complete the chapter.

The fifth chapter includes the AM-FM decomposition algorithm description and the glottal phases that are related to the AM and FM components of speech signal. In the last section of the chapter the observations of the AM and FM components during closed and open phases are discussed.

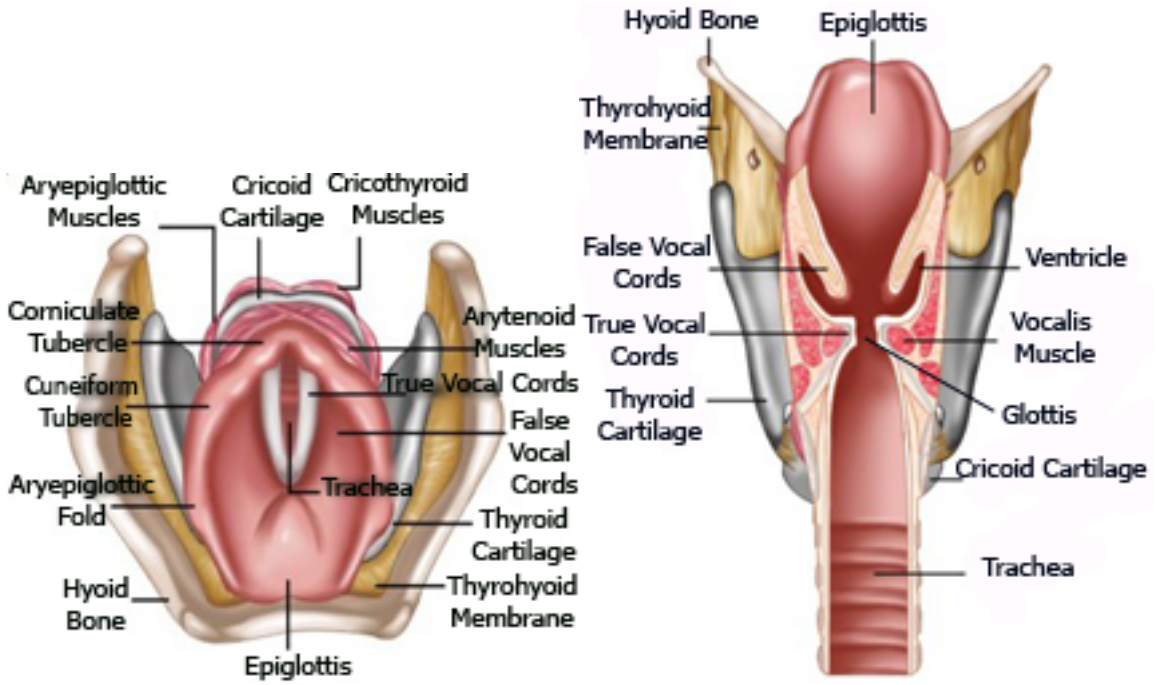
# Chapter 2

# Background

In this chapter, we examine the larynx physiology and the vocal folds' function, as it is beneficial in order to understand in depth the significance of electroglottographic signal and what it represents. The following section provides a description of the interaction of glottis with vocal tract though a developed model. The next section outlines the methods of vibratory cycle's recording. The fourth section of this chapter defines the electroglottographic signal and delineates the appliance of laryngograph. The phases and instants of an EGG signal and the corresponded intervals in glottal airflow are illuminated. Afterwards, the EGG signal is described in detail especially with respect to the shape of the waveform and to the time domain characteristics of the physiological features.

## 2.1 Larynx physiology and Vocal Folds

The phonatory process is produced while the air expelled from the lungs effects the vibration of the vocal cords, the glottis closes and the air flow is prevented. This glottal closure instant (GCI) is followed by an interval during which the glottis remains closed, until muscle tension and air pressure causes the folds to reopen at the glottal opening instant (GOI) [3]. The process is repeated periodically as a series of pulses that produces "modal" voiced speech. All voiced sounds are produced by an excitation signal that is filtered by a passive resonator called the vocal tract. This excitation is produced by the vocal folds that form a constriction at the top of the trachea as it joins the lower vocal tract. The glottis as we can see in Figure 2.1b covers the vocal folds and the space between the folds. The epiglottis, which is shown at the top of the figure shaped as a leaf, causes the intervention of the cartilage directing the food and liquid into the esophagus and protecting the vocal folds and airway during swallowing. A superior view of the vocal folds is shown in Figure 2.1a. The larynx's cavity consists of two pairs of soft tissue folds. The inferior folds are called vocal or true cords and above of them are the vestibular or ventricular or false cords [21]. The false folds do not usually vibrate with voicing, but they often adduct each other in people with muscle tension dysphonia, a disorder defined by excessive muscular tension with voice production. On the other hand, the true




(a) Larynx view from front

(b) Larynx view from top

Figure 2.1: Larynx physiology

vocal folds open, while we are breathing and close during voicing, coughing and swallowing.

The glottis controls the airflow towards the vocal tract. In voiced speech, the primary acoustic excitation normally occurs at the instant of vocal-fold closure. This marks the start of the closed-phase interval during which there is little or no airflow through the glottis. Acoustic theory shows that, for vowel sounds, the vocal tract acts as an all-pole filter whose input is the volume velocity of air through the glottis. Below, in Figure 2.2 we can observe the glottal flow and its derivative [23]. The glottal phases figured are:

- The duration from *t1* to *t2* corresponds to opening phase
- Closing phase occurs the time interval between *t2* and *t3*
- Closed phase during *t3* and *t5*
- The return phase occurs from *t3* to *t4*
- Variable *To* refers to the time duration of a vibratory cycle
- Variable *Oq* describes the Open quotient that is assigned to the open phase time duration divided by the fundamental period
- Variable *Qa* represents the return phase quotient that is defined as the ratio between the effective return phase duration and the closed phase duration $(1 - \mathrm{Oq})\mathrm{T0}$. In the case of an abrupt closure of the vocal folds, $\mathrm{Qa} = 0$

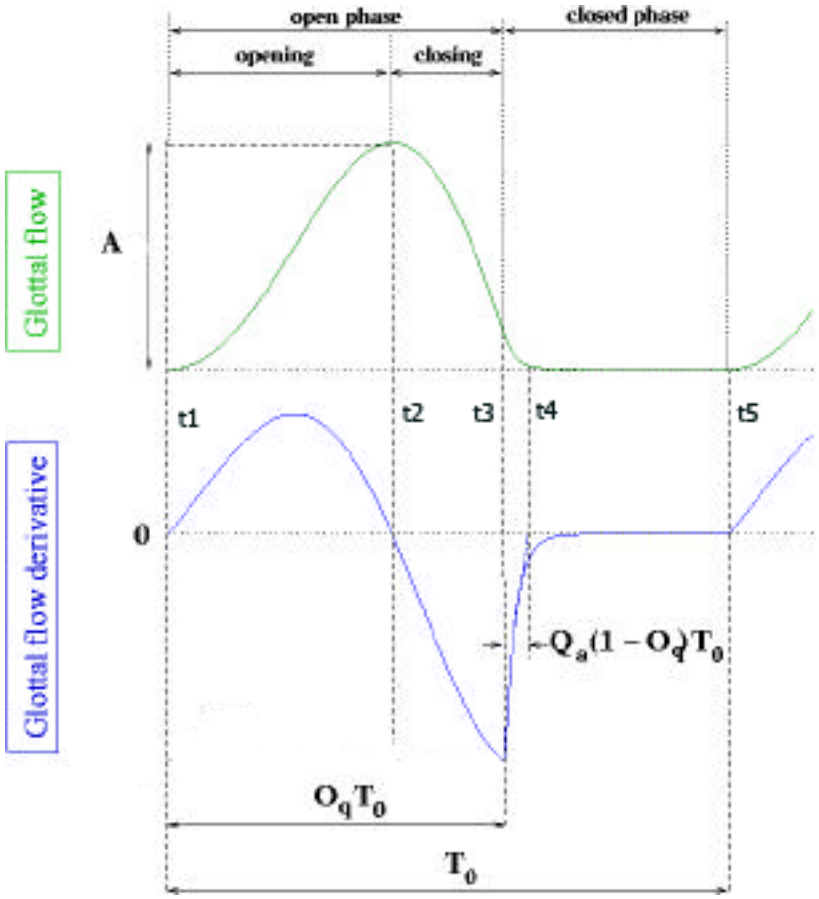



Figure 2.2: Glottal airflow and its derivative

## 2.2 Interaction of Glottis with Vocal Tract

A discrete–time model of speech production is developed in chapter 4 in [41]. The model was derived from a set of concatenated lossless tubes approximating the spatially–varying vocal tract, with a glottal input given by an ideal volume velocity source. Three speech sound sources were utilized: speech sounds with periodic, noise and impulsive inputs. This "source/filter" model initially assumed that the glottal impedance is infinite and the glottal airflow source is not influenced by the vocal tract. However, in reality a *nonlinear* coupling exists between the glottal airflow velocity and the pressure within the vocal tract. The pressure in the vocal tract cavity just above the glottis falls against the glottal flow and interacts nonlinearly with the flow. The vocal tract impulse response can abruptly decay within a glottal cycle. Thus, a simplified model of vocal fold/vocal tract coupling was used, which predicts this "truncation" effect and follows the modulation of formant frequencies and bandwidths within a glottal cycle. The formant frequency is proportional to the derivative of the area function and rises at the onset of the glottal open phase and falls near the termination of this phase. The opening of the glottis only increases the bandwidth of the formant.

Figure 2.3 illustrates the frequency and bandwidth of the time–varying first formant $F_1$

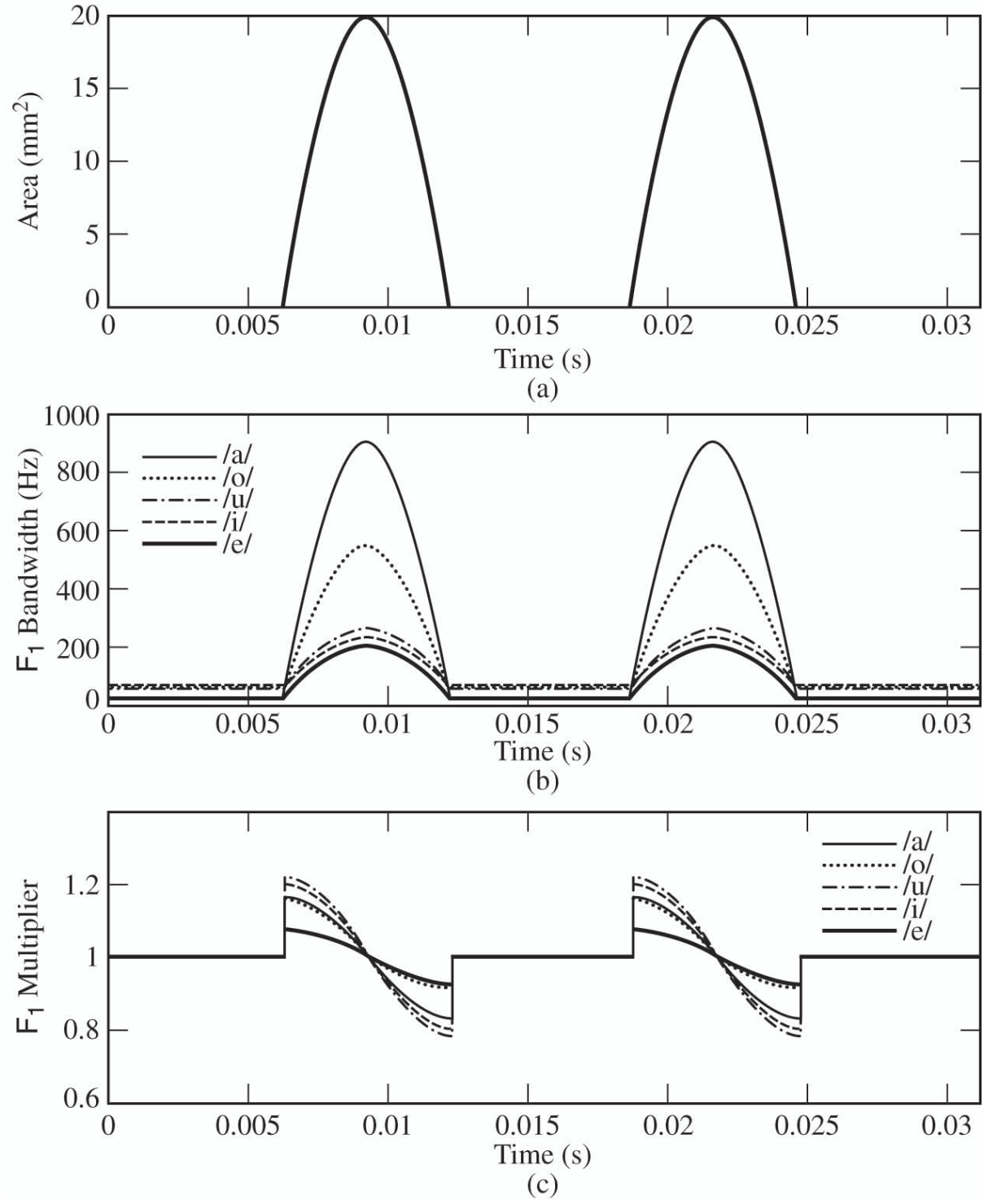



Figure 2.3: Time–varying first formant $F_1$ frequency and bandwidth for five vowels: (a) glottal area function, (b) bandwidth, (c) formant frequency

for five Russian vowels /a/, /o/, /u/, /i/, and /e/. The time–varying bandwidth (Figure 2.3b) is calculated as a function of time over two glottal cycles for the glottal area function (Figure 2.3a). The increase in bandwidth due to the opening glottis is higher for vowels with higher first formant. The change in formant frequency is rather instantaneous in both glottal opening and closure and can range between 10–20% of the formant frequency. A decrease in the impedance at the glottis as the glottis opens causes the "truncation" and the increase of the bandwidth within a glottal cycle. The "truncation" in a speech waveform is depicted in Figure 2.4. In the frequency domain, the energy drop, along with poor frequency resolution of the wideband spectrogram, prevents a clear view of formant movement. The wideband spectrogram (Figure 2.4b) was generated using a 4–ms analysis window and a 0.5–ms frame interval.

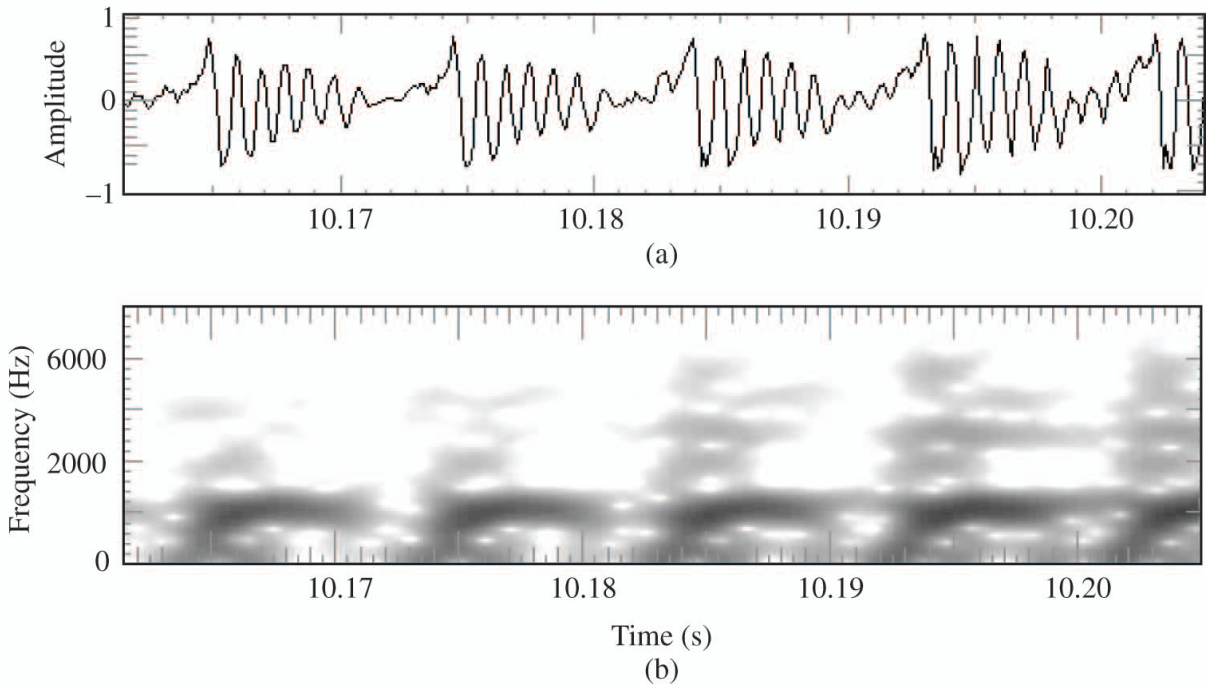



Figure 2.4: Truncation in a speech waveform: (a) segment from /o/ in "pop", (b) wideband spectrogram of (a)

## 2.3 Methods of Vibratory Cycle's recording

Alternative devices with varied procedures of recording the movement of vocal cords have been manufactured. The present techniques supported by separate devices are:

- Stroboscopy is a method of laryngeal endoscopy that depicts details for a precise diagnosis and assessment of vocal folds' flexibility. It is performed through a telescope or nasopharyngoscope. The obtained information concerns the nature of vibration and provides an immediate image to detect the presence or absence of vocal pathology and a permanent video record of the examination.

- High-Speed Stroboscopy, in which the voice source is sampled at very high rates, ranging from 1000 to over 8000 frames per second, as compared to the rate of 25-30 frames per second that is seen with conventional stroboscopic systems. It has been used in the quantification of normal and abnormal glottal vibratory patterns.

- Videokymography (VKG) was developed in 1994 in Groningen as a low-cost alternative to a high-speed imaging system, and is especially addressed for examination of vocal fold vibration. The system can work in two different modes: Standard and high-speed. In the standard mode, the VKG camera works like a standard commercial video camera with an image rate of 50 (interlaced) fields per second. In the high-speed mode, the images are recorded from a single line (selected from the whole video field) at a rate of around 7800 line images per second. However, this high-speed image rate is achieved at the expense of greatly reduced spatial information and this is a big drawback in VKG.

- Photoglottography or Photoelectric glottography (PGG) is a method that converts light intensity to electrical voltage. A light receptor is positioned at the glottis and while the glottis is open, more light is transmitted. This transmission is reflected in the waveform. Photoglottography assesses merely the open phase of the vocal cords.

- Electroglottography (EGG) becomes more and more popular as it is noninvasive, inexpensive, and easy to perform. An electrode is positioned on each side of the thyroid cartilage and a current passes through them. The result is a waveform that depicts the electrical admittance or impedance (depends on the device) of vocal folds during the current transmission. Further details of this recording device follow in the rest of the chapter.

- Ultrasound uses high frequency sound waves passed through body tissues. The sound waves are reflected at the interface between two media differing in specific acoustic impedance. The difference in acoustic impedance between tissues and the surrounding air is so large that the transmission of ultrasound from the tissue to air is negligible. In ultrasonography, two transducers are placed on both sides of the neck and echoes from the vocal cords are recorded. Images are obtained in rapid sequence and the appearance of continuous motion is given. Ultrasound can be a useful modality in the assessment of vocal fold vibration.

## 2.4 ElectroGlottoGraphic Signal

The non-invasive examination of vocal fold vibration has been performed by a simple electrical method described by the term 'ElectroGlottoGraph' or EGG for short. The device transmits a high frequency electrical current ($F \approx 1MHz$) of small voltage and amperage that passes through the neck of the subject. Between two electrodes the electrical admittance varies according with the vibratory movements of the vocal folds, while speech is produced, increasing as the vocal folds come in contact. The EGG signal is generated but also a simultaneous recording of acoustic signal is executed by a microphone producing a two-channel waveform.

Figure 2.5 indicates the apposition of the electrodes to the neck. The current passes through the skin of the throat and is detected by the electrodes. The received signal reflects the change in tissue admittance of the current's path. Afterwards, the signal is demodulated by a signal detector circuit, high-pass filtered, A/D converted and digitally stored. Figure 2.6 depicts the principle of the EGG device. The devices are commercially produced by Laryngograph Ltd., Synchrovoice, F-J Electronics, EG2-PC by Tigers DRS and Kaypentax Electroglottograph Model 6103. In this thesis the EGG signals needed for experiments are generated by Laryngograph Ltd, an admittance–based device [20, 18].

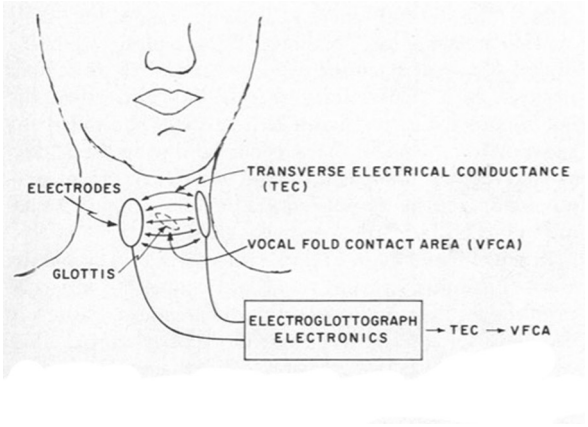



Figure 2.5: The apposition of the electrodes to the neck and the Vocal Fold Contact Area (VFCA)

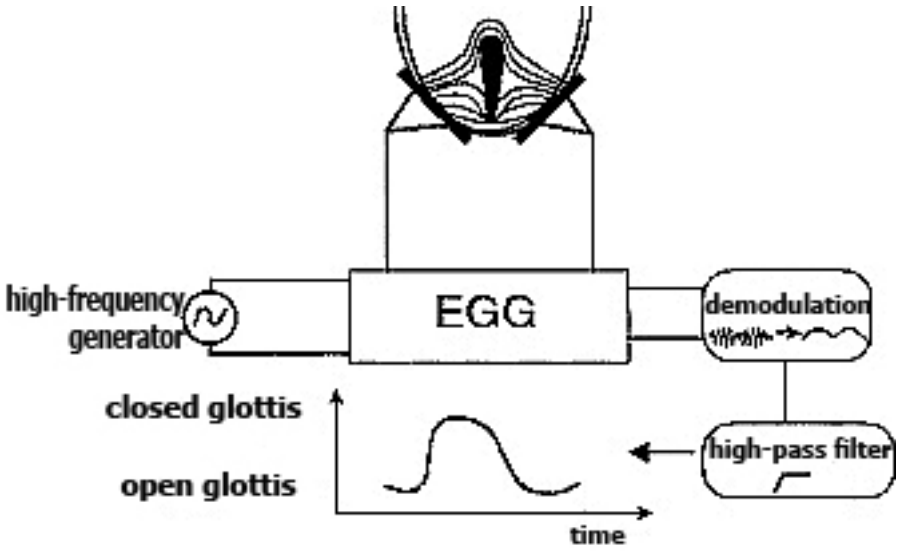



Figure 2.6: The principle of EGG device

Fabre (1957) was the first that applied the technique of Electroglottography. He introduced the term 'Electroglottography' (Electroglottographie) as the information derived from the device was related to glottal activity. Afterwards the term 'Laryngograph' appeared with the homonym device and the generated waveform was named 'Lx'.

The signal extracted from the EGG device contains information concerning the vibration of the vocal folds but also the slower movements of the other structures in larynx. Therefore, Fourcin and Abberton [19, 16, 17, 18, 1] proposed the name 'Gx' for the waveform of larynx movement and the name 'Lx' for the vibration component. The Gx component can be caused by swallowing and generally by the vertical movement of the larynx. Fluctuations of this type are usually slightly informative and are removed from further analysis. The effects of the varying larynx height are compensated by the use of additional electrodes or high pass filtering of the registered signal. These methods may involve signal distortion, especially for low-pitched voices. The distortions may be caused by a too high cutoff frequency of the filter (or a too wide filter transition band). This can cause the attenuation of the Lx signal component. The non-uniform phase response function of the filter may change the shape of the filtered waveform. Nevertheless, even the unfiltered output of the EGG device is not free of distortion. Particularly the demodulation circuit whose frequency transfer function may influence the frequency response of the EGG device, especially in the low frequency range, constitutes an additional source of signal shape deformation. Thus, many of the commercially available devices include an automatic gain control facility that is used to compensate for the variations of the signal level that are due to varying throat admittance. Such circuits respond with a small time delay and the time constant of the device may influence the EGG waveform at low frequencies.

The proper placement of the electrodes is of high importance since a slight shift might cause spurious effects in the recorded signal. The generation and the amplitude of the EGG signal depends on:

- the configuration and placement of the electrodes, the signal to noise ratio (SNR) is optimized when the electrodes are positioned at the level of the vocal folds
- the distance between the electrodes
- the electrical contact between the electrodes and the skin
- the position of the larynx and the vocal folds. Vertical larynx height changes for different articulations and phonational qualities (*F0*). This results in a change in the relationship of the electrodes to the vocal folds and thereby influences the EGG waveform.
- the amount and proportion of muscular, glandular and fatty tissue around the larynx. Fat tissue is a very poor conductor. A fatty layer under the skin can degrade the LX signal considerably.

During a vocal-fold vibratory cycle, the corresponding EGG signal can be described by four main phases, as it is illustrated in Figure 2.7:

- 1-3: Closing phase The lower margins of the vocal folds initiate the contact *(1-2)* and then upper margins are coming in contact as well *(2-3)*. As closing is generally faster than opening, this phase is characterized by a steep slope in the EGG signal. The instant of maximum slope can be found at *2*, which corresponds to a strong positive peak in the derivative of the EGG signal "DEGG" signal.

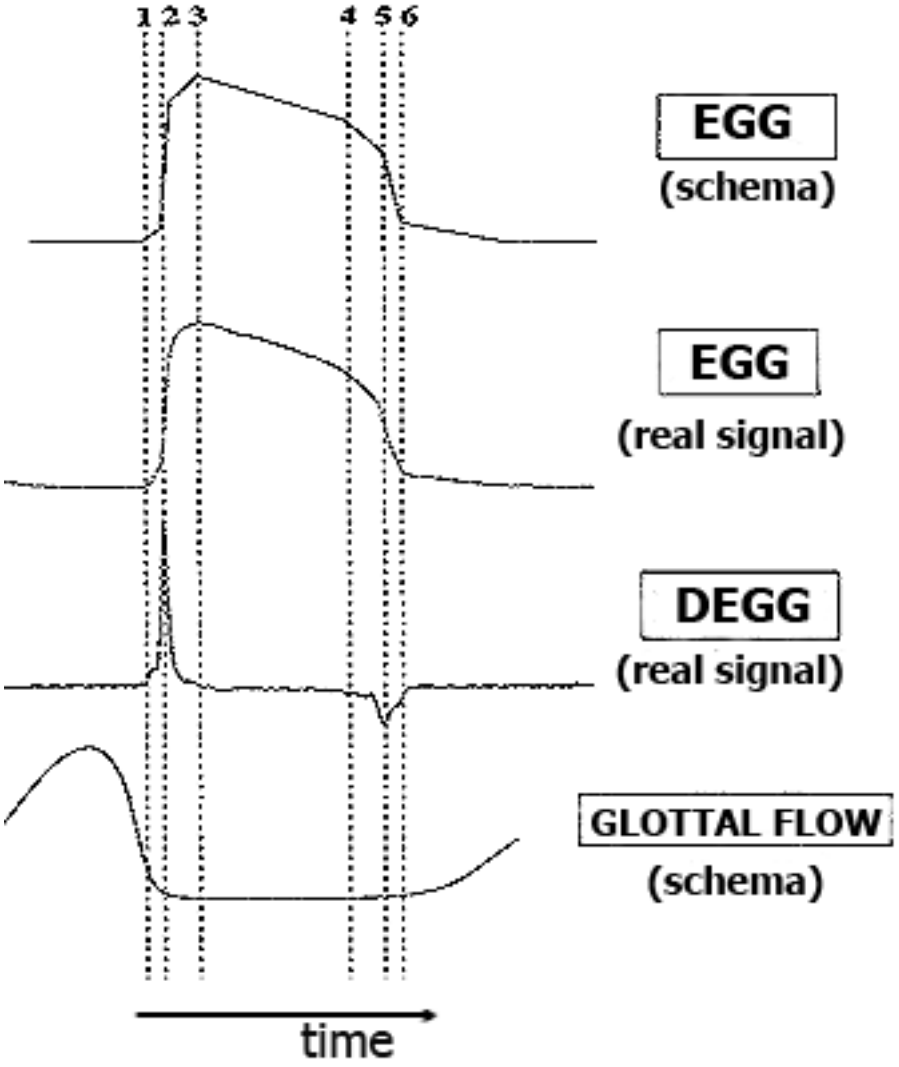



Figure 2.7: The vibratory cycle depicted in EGG signal and Glottal Flow

- 3-4: Closed phase At this interval the vocal folds are in full contact, preventing air from passing through the glottis.
- 4-6: Opening phase The separation of vocal folds begins from the lower margins *(4-5)*, followed by separation along the upper margins *(5-6)* gradually. The negative peak in the "DEGG" signal that corresponds to the instant of maximum slope *(5)* is weak due to the gradual separation.
- 6-1: Open phase The vocal folds are in no contact. A relatively flat signal is observed, as there is little variation in the admittance.

An example of a speech waveform and the respective EGG signal with GCIs' and GOIs' markers are shown in Figure 2.8. A series of eight stroboscopically derived images of the vocal folds with the corresponding EGG waveform under each image is pictured in Figure 2.9. Each image represents a different vocal folds' phase:

- Case a: Closure Instant (GCI)
- Cases b, c: Instants in Closed phase
- Case d: Opening Instant (GOI)
- Cases e, f, g, h: Instants in Open phase

If the vocal folds adduct very rapidly and all along the vertical movement, the closing and closed phases become indistinguishable and consequently the slope of the closure phase becomes steep. The slope of closing segment is more gradual than the slope of closed in low to normal recorded intensities.

The time derivative of the Lx waveform is widely processed in the detection of signal's periodicity. It is also beneficial in identifying the distinguishable changes in the slopes during the phases of increasing and decreasing current-admittance of vocal folds. The positive peak of the derivative is regarded as an indicator of the GCI. The computation of fundamental frequency from an acoustic signal is more complicated and less precise than extracting it from the derivative. Furthermore, if a relatively strong EGG is registered, it is preferable to refer to the derivative for an accurate measurement of speech fundamental frequency (*F0*). However the signal-to-noise ratio (SNR) should be low and *F0* should not be measured during voice offset and onset, whereas several EGG pulses can exhibit a distorted shape, in order to obtain valid results. Pitch or Fundamental period is usually defined as the duration between maximum positive peaks in the differentiated EGG waveform. The inverse of Fundamental period gives the Fundamental frequency of the voice. The marking of pitch period is usually done by algorithms that use a threshold value to detect the peaks of the signal derivative. The threshold is usually defined as a medium value between the minimum and the following maximum peak of the waveform. Nevertheless, this method does not always brings reliable results even for normal voices due to the rapid baseline changes caused by the vertical larynx

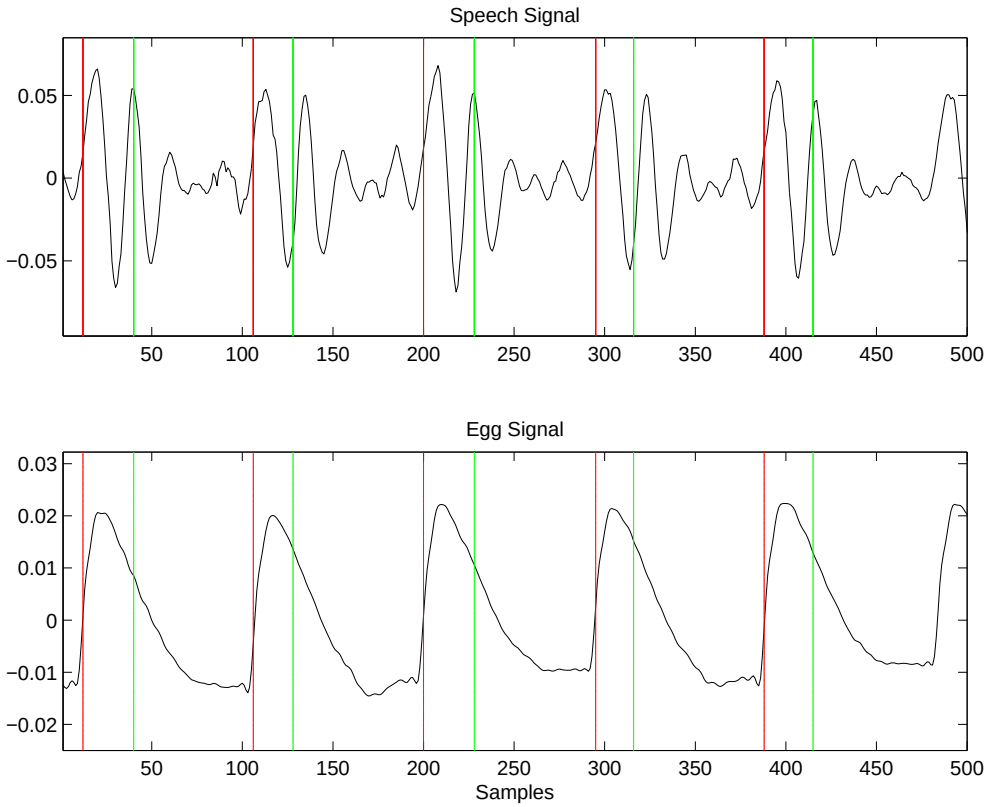



Figure 2.8: Speech and EGG waveform from a female young speaker with GCIs and GOIs markers from the sustained vowel /a/. (GCIs-Red Color, GOIs-Green Color)

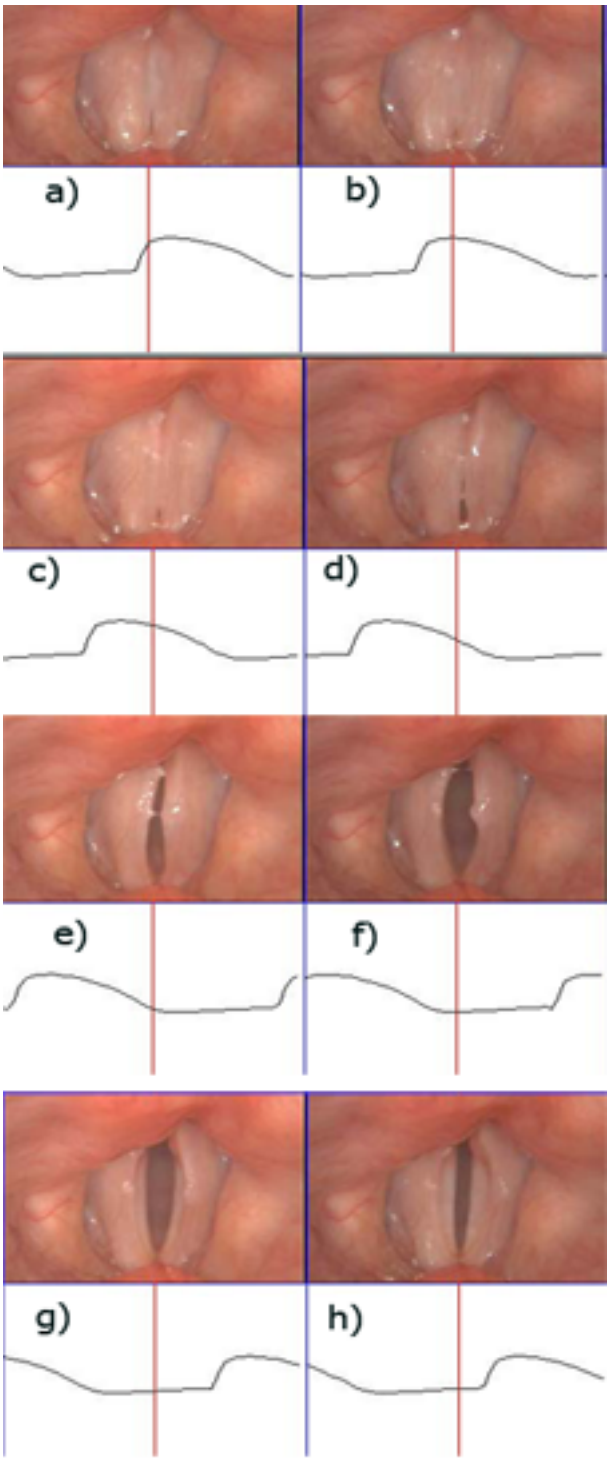



Figure 2.9: Stroboscopy and EGG aligned waveform in modal voice

movements (Gx signal). In these cases multiple peaks in the DEGG signals may occur even for modal speakers or if the signal is noisy.

On the other hand, the negative peak of the EGG derivative indicates the GOI. The EGG waveform at the Opening Instant descents and the minimum of the derivative's time waveform corresponds to that moment. Firstly, the EGG decreases monotonically, reflecting the decrease in lateral contact between the vocal folds. During this interval the EGG waveform is convex as it is illustrated in Figure 2.8. Then, as the upper margin of the vocal folds separates, the waveform of the EGG changes to a concave shape. There have been remarkable observations that add doubts about this procedure of GOI's detection supporting that there is an impact of additional mechanisms that can mask or delay this moment in the EGG signal. There is evidence for the mucus bridging effect, in which a strand of mucus can bridge the glottis as the vocal folds initially open. When the mucus bridge breaks, the EGG waveform records a sharp fall even though the glottis is already open.

GOI is defined equally to the *Open Quotient* (Oq) of the glottal flow provided the assumption that the opening instant is determined correctly. The duration between the glottal opening instant and the consecutive glottal closing instant corresponds to the open time. The Open Quotient can then be derived from these two measures as the ratio between open time and fundamental period. The Open Quotient indicates the duty ratio of the glottal airflow. A change of the duty ratio substantially changes the spectrum of an excitation. Additionally, it is highly correlated to physiological constraints, as is the case in different phonation types. For instance, the breathy voice phonation quality is characterized primarily by its longer open phase [29]. The *Closed Quotient* (Cq) is respectively defined as the ratio between the time duration that vocal folds remain closed and the glottal period. Open and Closed Quotient are of significant value as they comprise a voice quality measure. Changes in inferred vocal fold open/closed quotient are considered as an aspect of acoustic efficiency variation. In experiments related to voice pathology any of the quotients are examined for the pathology evaluation [12]. Furthermore, in case that a therapy is addressed, the Cq is measured before and after the treatment. The majority of voice pathology demonstrate a low closed quotient and consequently, a high open ratio. After the received therapy the closed ratio is usually increased. The quantification of Cq expedites the professional singers. The trained singer by increasing the ratio of vocal folds' contact is able to make use of a natural acoustic consequence of an adjustment to the manner in which the folds vibrate to increase overall system efficiency. A different feature was studied in [38]; the "Derivative EGG Closure Peak Amplitude" (DECPA), the amplitude of DEGG signal at glottal closure instant. The study discusses the correlation between the DECPA and the accent

The admittance fluctuation caused by the vocal cords' movements is too weak to be registered for many speakers. EGG signals of acceptable quality are harder to obtain from women and children than from men. This is related to the smaller mass of the vocal folds, the wider angle of the thyroid cartilage and different proportions between various types of tissue. In the case of pathological voices the location of the closure instant in the EGG signal is no longer as obvious as for normal voices. The adduction phase is not always smooth and additional peaks in the DEGG waveform are often observed. In breathy voice the vocal tract resonances are less well defined in amplitude because this voice quality is associated

with a less rapid closing phase [36]. Although Lx waveform seems periodic, breathy voice gives an auditorily perceptible component of irregularity in open intervals of the speech signal waveform from period to period. The EGG waveform does not reliably reflect the glottal activity for voices with a continuously open glottis. In that case, the variation of the larynx admittance does not correspond to the glottal area. Particularly the amplitude of the Lx waveform fluctuation may change in accordance with the reduced contact between the folds.

At last Electroglottography provides a better representation of the closed and closing phases, especially of the vertical contact area, than the other relative methods. The high advantage of EGG relies on the fact that it is not uncomfortable to speakers, as it does not influence at all the articulation and voice production. In that way it is considered to be non-invasive.

# Chapter 3

# Related Work

Much research work has been conducted on the detection of glottal instants from the EGG signal and these studies are included in the first part of this chapter. The next part is devoted to the derivative of EGG, which has not attracted much attention until the past decade, although it yields reliable indicators of GCIs. The last part of the chapter discusses the extraction of the glottal Instants from the speech signal that has drawn widely the interest with various approaches.

## 3.1 Determination of Glottal Instants in ElectroGlottoGraphic Signal

Several researches have analyzed the EGG signal and compared it with other methods such as stroboscopic photography, high–speed cinematography, photoglottography [13]. All of these studies have generated waveforms from the mentioned techniques with simultaneous recordings with EGG and compared the results in order to evaluate the Electroglottographic signal. The examined features were the GOIs, GCIs, the instant of the maximum opening of the glottis, open quotient and the relative average perturbation measured from the glottal area to that estimated from the EGG. The experiments indicated that the vocal–fold contact area (VFCA) was reflected to the EGG waveform, as the larger the contact surface, the larger the measured admittance.

The detection of glottal Instants has mainly based on the EGG signal. The common applied method is the **threshold–based** that has been widely used and it is shown in Figure 3.1 [42]. The threshold or criterion–level (hereafter CL) can be determined in few ways, with the most apparent methods involving a comparison of either the area (time–integral) or the amplitude of the waveform above and below a proposed CL. The definition of a CL is equivalent to making the CL the average value of the waveform during each cycle. The selection of a CL from the maximum and minimum amplitudes of the waveform during the vibratory cycle will be referred as using a %–level criterion or threshold. The easiest aplied

Criterion–Level is the average–value level (50%), however, the threshold 35% reflects better any existent greater degree of vocal fold adduction. Another threshold that has been initially tested in open quotient estimation is 70%.

Generally, the threshold–based technique in EGG involves the selection of a level line as a percentage of the amplitude between the minimum and the maximum of the signal over a glottal period. The two crossing points between the level line and the EGG signal are considered as the glottal closure and opening instants respectively. These points are used for the open (or closed) quotient approximation. This kind of thresholds is very convenient for medical purposes, as it can be applied even on noisy or weak signals. However, the results of such methods are a priori unprecise compared to that measured on a glottal flow signal. The major drawback of such methods is the lack of accuracy for the GCI detection and missing of GOI.


Figure 3.1: Illustration of three threshold–based methods for measuring the open quotient (or its equivalent, the closed quotient) on an EGG signal

## 3.2 Observations and Determination of Glottal Instants in Differentiated EGG

### 3.2.1 Observations in DEGG

The hypothesis that DEGG signal can provide us with reliable indicators of Glottal Instants has been supported with simultaneous recordings of EGG, inverse–filtered derived glottal flow and ultrahigh–speed cinematography. The signals were synchronized in order to be evaluated. In non–pathological cases, it was proved that the observed peaks in DEGG were related to the GCIs and GOIs, defined as the instants of initialization and termination of glottal area variation. Other simultaneous measurements with electroglottograph and photoglottograph pointed out that DEGG peaks were related to the peaks observed in the derivative of photoglottographic signals and to the opening and closing instants extracted from the glottal flow and its derivative.

The peaks in DEGG are not always well–defined [22]. Some cases are shown in Figures 3.2, 3.3, 3.4. The first case (Figure 3.2) indicates the absence or imprecision of the opening peak. The DEGG signal in the left column has no negative peak to be detected, thus an opening instant is absent. More than one negative peaks are observed in the right DEGG signal and one of them should be pointed as GOI.

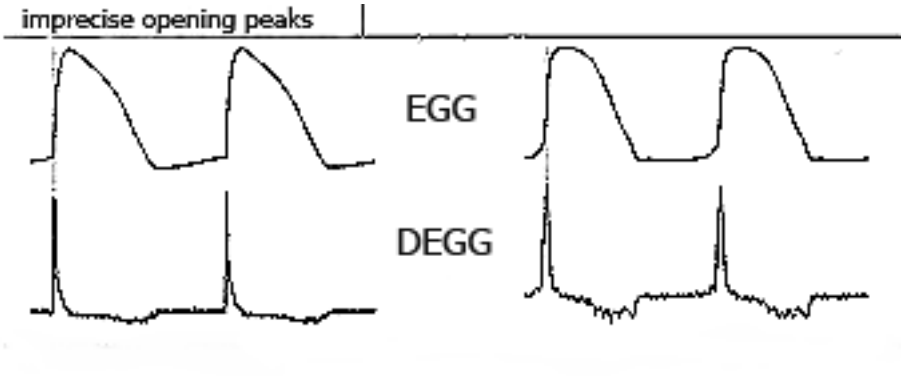



Figure 3.2: Examples of two–period EGG and DEGG signals, where the opening peaks are imprecise

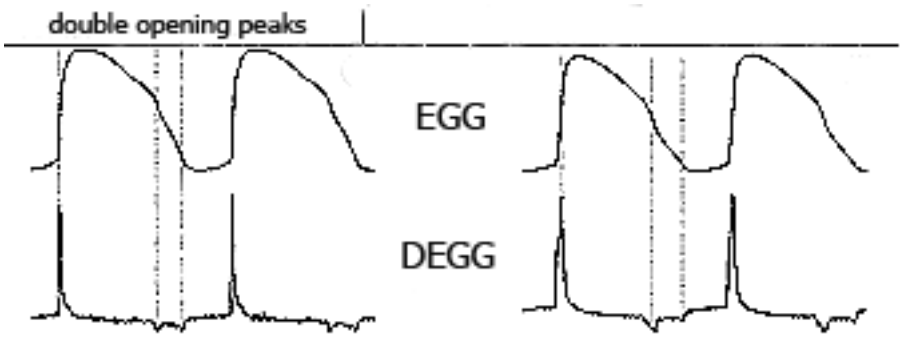



Figure 3.3: Examples of two–period EGG and DEGG signals, where the opening peaks are double

Although the closing peak is precise, in several recordings two peaks are observed as illustrated in Figure 3.4. The opening peaks may be double as it is shown in Figure 3.3. The reported double–peak opening (or closing) feature is related to either a difference in how opening (or closing) takes place over the thickness of the vocal folds or to a time–lag opening (or closing) at two different parts of the glottis. A time–lag may occur due to the time difference between glottis–anterior opening (or closing) and glottis–posterior opening (or closing). Conclusively, the open quotient and glottal Instants' determination cannot be accurately measured on a DEGG signal, unless the closing and opening peaks are single and precise.

### 3.2.2 Determination of Glottal Instants in DEGG

The EGG signal is proportional to the glottal contact area, whose derivative during voiced speech is an impulse train. Many approaches exploit this property and they are discussed below.

**Threshold–based methods**

The dominant algorithms applied for the determination of Glottal Instants are the threshold–based ones. The positive peak in DEGG indicates the GCI and the negative one the GOI. Different thresholds are used to detect peaks related to the maximum value (for GCIs) and the minimum value (for GOIs). The common implemented threshold has been 50% for each peak and it is illustrated in Figure 3.5. Nevertheless, unless the peaks are both precise, the Instants and the open/closed quotient is difficult to be measured.

These approaches make use of dynamic thresholds to obtain an accurate estimation of GCI during voiced speech. Howard [26, 25] proposed a method, where the GCI is detected from the DEGG positive peak and the GOI is estimated through an EGG–based threshold method. The GOI detection should be more accurate than previous methods due to its connection to the respective GCI. The EGG signal with a 3/7 threshold difference between the minimum and maximum value of the signal over a glottal period is used in combination with the detection of glottal closing instants in the DEGG signal. The article, which suggests this method, describes a measure that ranks trained and untrained singers appropriately. The hypothesis of the study was that during singing, the laryngographically derived Open Quotient is reduced with experience and training. The vocal improvement related to Open and Closed quotient measurements from electrolaryngograph were compared to the measurements from an automatic inverse filtering technique. The derivation of Closed quotient (Cq) from the electroglottographic output that applied in trained and untrained female singers is depicted in Figure 3.7.

The subsequent applications use a threshold in DEGG for the location of glottal Instants. SPAR (Speech–Pattern Algorithms and Representations)[28] (1987) is a project concerned with advanced speech analysis algorithms that provides a unifying framework for software development and a consistent user interface. Many different representations of a speech token can be stored in a single speech file. A speech file may contain an unlimited combination and repetition of types, such as Speech pressure waveform, Laryngograph waveform, Fundamental frequency, Points of voiced excitation and others.

SPAR was improved and the SFS (Speech Filing System) [27] (2008) was recently developed. The algorithm *txgen* in SFS generates open and close phase markers from laryngographic signal. This algorithm consists of the following steps:

- *Low–pass filter the Lx waveform at 3000Hz*
- *Differentiate the Lx waveform*
- *Find maxima and minima of differentiated signal that lie above a threshold. These are taken as the positions of closure and opening on the Lx waveform.*

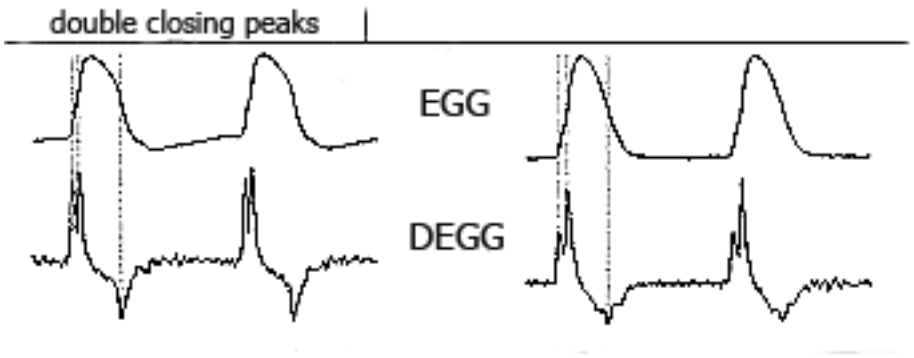



Figure 3.4: Examples of two–period EGG and DEGG signals, where the closing peaks are double

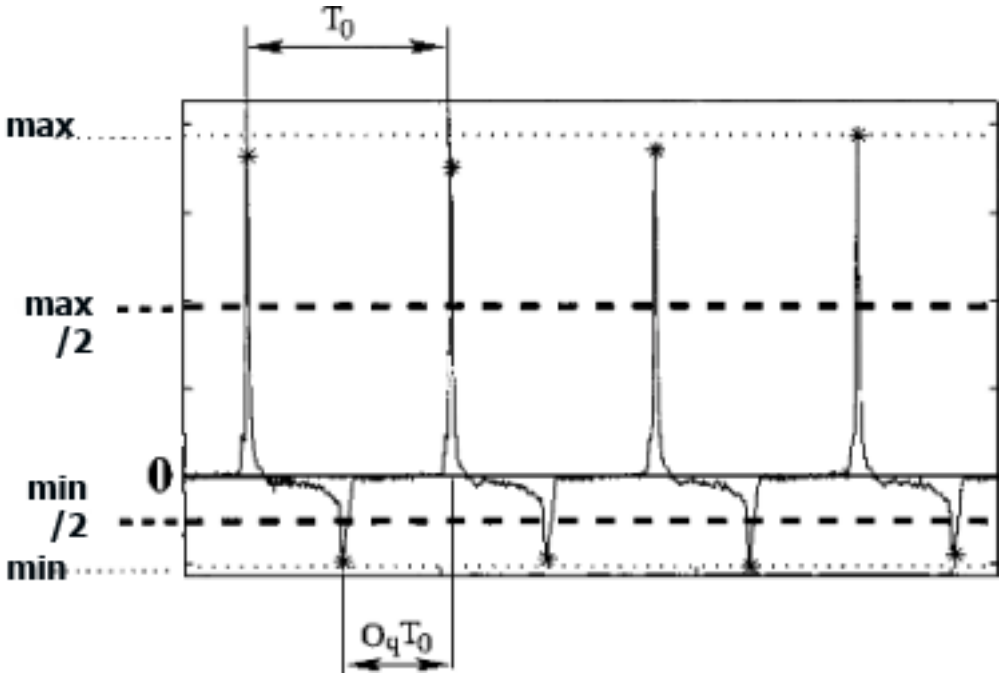



Figure 3.5: Illustration of a threshold–based method for measuring the open quotient (or its equivalent, the closed quotient) on a DEGG signal

- *The positions of larynx opening are determined as the next point in the Lx waveform at amplitude equal to that at closure*

The function *High–Quality Tx (HQTx)* produces excitation periods from a laryngographic signal. The characteristic shape of the Lx waveform during voiced excitation is exploited. Two derived functions are utilized: the sample–by–sample differential and the instantaneous gradient. Tx points are extracted from these functions using an automatic threshold determination procedure.

The thresholding methods applied in simple derivative lack in GOIs identification, since the derivative shows no clear negative peaks in majority of real recordings. GOIs are principally associated to their distance from GCIs and they are not detected separately. The SFS described earlier shows several GCIs within a glottal cycle, although only single positive peaks exist in DEGG signal. GOIs are misidentified, since they are related to misdetermined GCIs.

**DECOM**

A different DEGG–based method DECOM (DEGG Correlation–based method for Open quotient Measurement) was proposed by Henrich[22] in 2004. DECOM can be applied to the case of quasisteady voiced sounds and it is based on the standard methods for fundamental frequency estimation of a voiced signal. The principle of the DECOM method is to measure fundamental frequency ($F0$) and open quotient ($Oq$) using a correlation based method to estimate the distance between two consecutive closing peaks and the distance between an opening peak and the consecutive closing peak. In addition, an automatic detection of the number of peaks at closing (npeak closing) and at opening (npeak opening) is performed. The derivation of DECOM's open quotient is compared to EGG–based threshold techniques and to inversed–filtered glottal flow measurements. The open quotient measurements using the DECOM method are, on average, in much better agreement with the glottal–flow measurements than most threshold–based methods for sustained phonation, when double peaks are rejected. The corresponding algorithm with a schematic description is presented step–by–step in Figure 3.6.

The DECOM method focuses on open quotient measurement and double peak detection, exploiting the simple derivative. This method is more robust when double peaks exist in DEGG signal. The double peaks are determined and are excluded in open quotient computation.

**SIGMA**

The SIGMA (Singularity detection In EGG with Multiscale Analysis) algorithm [50, 51], introduced in 2007, describes an alternative method for EGG–based GCI estimation, which is based on the stationary wavelet transform. The GCI detection is executed with a group delay function and Gaussian Mixture Modeling is applied for discrimination between true and false GCI candidates. In this group delay approach the negative–going zero crossings of the average slope of the negative unwrapped phase of the Fourier transform of the EGG derivative are identified and calculated over a sliding

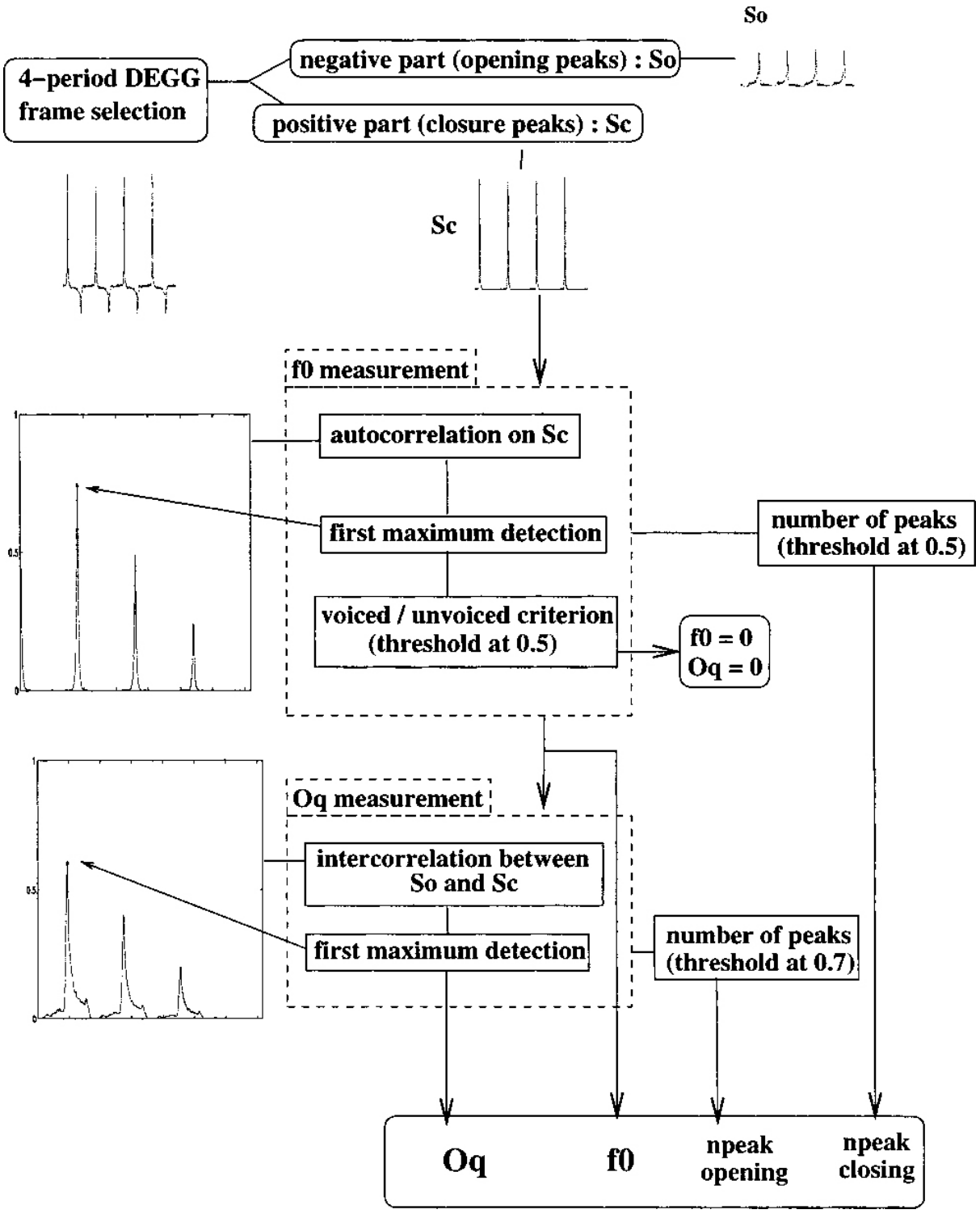



Figure 3.6: Schematic description of the DECOM algorithm

window. The number of false candidates that may arise are removed by modeling three–dimensional feature vectors as Gaussian distributions and clustering with an unsupervised EM algorithm. This method compares the generated GCIs to the respective ones of HQTx algorithm (SFS). The HQTx appears to be prone to detect more GCIs per glottal cycle than the correct ones.

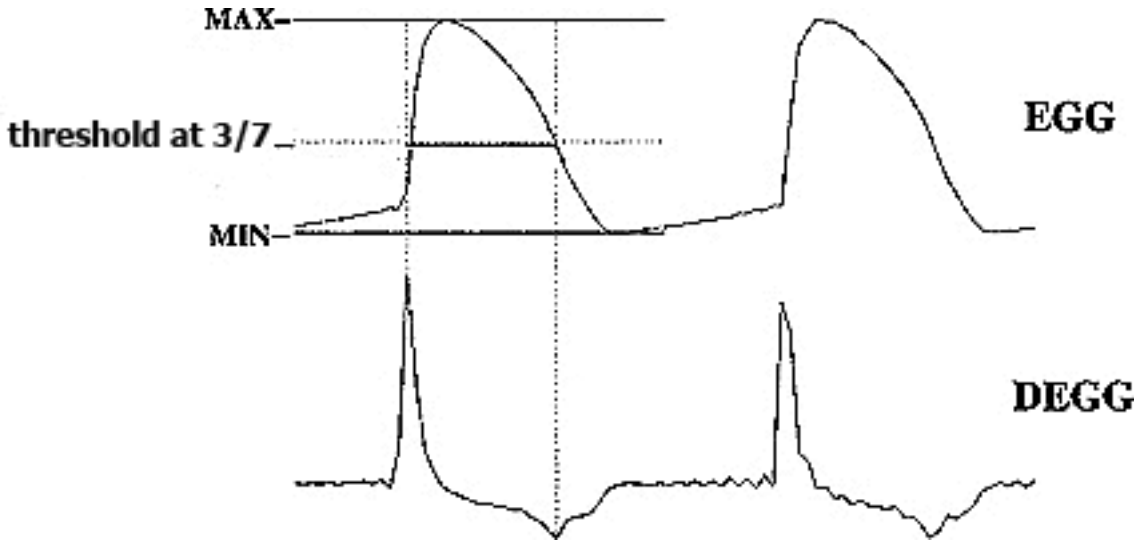



Figure 3.7: Illustration of a threshold–based method for measuring the open quotient (or its equivalent, the closed quotient) on an EGG signal and its derivative

## 3.3 Determination of Glottal Instants in Speech Signal

Remarkable research work has been conducted on the detection of glottal closure and opening instants from the speech waveform, without the use of the electroglottographic signal (EGG). The proposed algorithms may rely on speech signal's energy, linear predictor, and/or group delay measure. Therefore the following categories are not independent and few methods use a combination of them.

**Linear Prediction Model**

The vocal tract reflects a strong excitation when the glottis closes abruptly. Free oscillations follow the closure, since the glottis opens more slowly than the closing. Thus the glottal closure instant will lie close to the most noticeable increase of amplitude of the sound pressure. Consequently, the prediction error will be distinguishable there, since the signal cannot be well represented. The free oscillations render the prediction knowledge inadequate. Several ambiguities appear in the prediction error and there are no definite intervals where the GCI occurs.

Linear prediction (LP) of the sound pressure is a useful means for extracting various physical parameters of the speech signal for coding or analysis. Linear Predictive Coding (LPC) residual is obtained by applying the LPC inverse filter to a preemphasized speech waveform. Its waveform includes an impulsive feature at closure and a similar but smaller impulse at opening. The use of this LPC residual waveform for detecting glottal closure instants requires the following assumptions: (i) the vocal tract acts as an all–pole filter, (ii) the filter can be estimated adequately from the speech waveform alone and (iii) the LPC residual will contain an identifiable impulse at closure for voiced speech sounds.

Several approaches search for discontinuities in a linear model of the speech production. One of the earliest approaches applied a predictability measure to determine GCIs by finding the minimum of the Gram determinant of the autocovariance matrix of the speech signal [47] in 1974. The method was tested with stationary male vowels only and produces inadequate results for some vowel sounds. Furthermore, many covariance matrices and their determinants must be formed resulting in a computationally expensive algorithm. In 1979 [56] GCIs and GOIs were detected at the minimum energy in the LPC residual, using a sliding window. The approach is based on a sequential covariance analysis and a normalized total squared error criterion.

In [2], Ananthapadmanabha and Yegnanarayana implement the location of discontinuities from the derivative of the glottal air flow. A thorough analysis of LP residual indicated that the ambiguities arise mainly due to zeros in the vocal tract system and the phase angles of formants at the instant of excitation. The method relied on computing the Hilbert envelope of the error signal and was tested by extracting glottal pulses and comparing the instants of slope discontinuities in the glottal pulses with the epoch locations. The detection may be imprecise, since noise may produce similar discontinuities to those of the voice production. In 1999 [7] expressions for the flow of acoustic energy in the lossless–tube model of the vocal tract are suggested. It was shown that linear predictive analysis could be used to estimate the waveform of acoustic input power at the glottis and identify the glottal instants during voiced speech.

Subsequent LP analysis on the closed–phase speech data is executed in [37] in 2001. The closed phase of the glottal cycle is determined with the exclusion of intervals that are not within the closed phase. The measure used is the log determinant of the Kalman filter estimate error covariance matrix.

The recently proposed method [15] in 2009 firstly computes intervals where the glottal instants are expected to occur and then detects a discontinuity in the determined area of the linear prediction residual.

### Energy Peaks

An alternative approach of detecting glottal Closure Instant is to search for energy peaks in waveforms derived from the speech signal. As a development of [47], in 1994 [33] the GCI is identified as the maxima of the Frobenius Norm of the signal matrix and the Singular Value Decomposition (SVD) method is utilized. The Frobenius Norm offers a short–term energy estimation of the speech signal. An energy value is assigned to each speech sample with the usage of a sliding window. The computational efficiency of the algorithm is based on the calculation of the Frobenius norms of signal matrices.

Many approaches rely on energy peaks in waveforms derived from features in speech time-frequency representation. In [54] (1999) the wavelet transform used to represent the speech and determine glottal closure instants. The analysis is based on a dyadic wavelet filterbank. Lines of amplitude maxima in the time-frequency plane are defined using a dynamic programming algorithm. Lines seems strong for voiced speech and weak for unvoiced speech. The GCIs are identified to correspond to the line with the maximum accumulated amplitude within each pitch period.

Very few works have aimed in the determination of GOI from speech signal. The energy of the excitation at GOI is weaker and more dispersed than at GCI, rendering its detection rather difficult. In 2004 a method for detection of both glottal instants based on a multiscale product of wavelet transforms was proposed [4] with no quantitative results. Various scales are applied that produce the wavelet transform coefficients in order to enhance edge detection and estimation. A relevant approach is adopted in [5, 6].

**Group Delay**

A group delay function can be evaluated for either the speech signal or the LPC residual to detect GCIs. Several methods exploit the phase properties of GCI (impulse) by computing a group delay function.

The difference between the methods depends on what measure is used to determine the group delay of the sliding window. A measure applied [45] in 1995 is the average value over all frequencies. It was the first attempt of using a group delay measure to determine the acoustic excitation instants. The method calculates the frequency–averaged group delay over a sliding window applied to the LPC residual. The measures proposed by Stylianou [49] in 1999 are the zero frequency value of the group-delay and the energy–weighted phase. Both of them were applied to the problem of inter–segment coherence in concatenative speech synthesis. The last suggested measure in 2002 is the energy–weighted average of the group delay applied in DYPSA algorithm [30, 31, 39, 9, 34, 35].

The Dynamic Programming Projected Phase–Slope Algorithm (DYPSA) algorithm extended the technique of group delay. The projected phase–slope estimates GCIs candidates that previous algorithms missed. The phase–slope function is defined as the slope of the unwrapped phase of the short–time Fourier transform of the LPC residual. The GCIs are identified as positive–going zero–crossings in the phase–slope function. A dynamic programming (DP) with a cost function is employed to eliminate spurious detections that correspond to glottal openings or other events. The computational efficiency of the algorithm is achieved with the retainment of the N–best path segments at each stage of DP. In [9] it is proved that if the measures are applied to the preemphasized speech instead to the LPC residual, the timing accuracy worsens but the detection rate improves slightly. In [35] an improvement of DYPSA algorithm is presented. A voiced/unvoiced/silence discrimination measure, based on speech extracted features, is additionally applied in order to minimize the false candidates of GCIs. Speech classification is performed that classifies candidates as voiced or non–voiced. An application of the DYPSA algorithm to segmented time scale modification of speech is included in

[52]. The DYPSA method and Multi–Scale product are evaluated in [43](2008) and it is reported that Multi–Scale product detects more precisely GCIs.

**Other Methods**

The nonlinear components of speech signals are generated by several effects based on the speech production process and the excitation. An approach based on a weighted nonlinear prediction was proposed in [44] in 2007. Feature signals are obtained from the nonlinear prediction of speech using a sliding window technique. The glottal closure causes a maxima in the feature signals relative to that from DEGG signal. This contribution proved that small prediction orders and individual nonlinear coefficients are beneficial.

In 2009 [14] proposed a method of detecting GCIs using a glottal shape estimation and a standard lips radiation model. The speech spectrum divided by the vocal– tract filter generates the time–derivative glottal source. The GCIs are easily identified by the derivative of glottal source.

# Chapter 4

# Determination of Glottal Instants in EGG

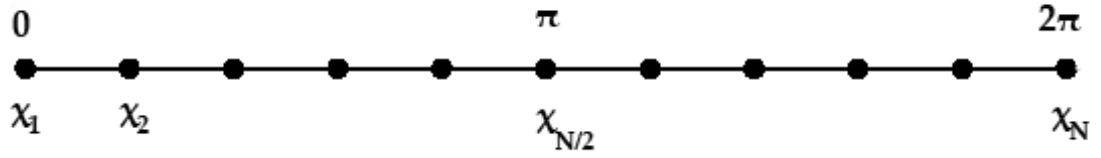
This chapter introduces an alternative method to the conventional derivative. This approach is based on the spectral methods [53] that are thoroughly examined. The spectral differentiation is introduced and its usage to the EGG signal. Afterwards, a new way of differentiating the EGG signal is presented. The method was proposed in [55] and is called "Slope Filtering". The filtered EGG signal provides a distinct illustration of its slope and the main glottal Instants are well distinguished. Furthermore, the results of the applied methods and the experimental database are explicated. The conclusions of the current approaches are stated at the end of the chapter.

## 4.1 Spectral Methods

The fundamental principle of spectral collocation methods, for a finite grid, is to interpolate the data globally and then evaluate the derivative of the interpolant on the grid:

- *Given $v$, its band–limited interpolant $p$ is determined*
- *Let $p$ be a single function (independent of $j$) such that $p(x_j) = u_j$ for all $j$.*
- *Set $w_j = p'(x_j)$.*

The proposed method focalizes on a spectral differentiation on a bounded, periodic grid. The basic periodic grid will be a subset of the interval $[0, 2\pi]$:




The translations to other intervals such as $[-\pi, \pi]$ exhibit the same behavior. The periodic grid implies that any data values on the grid are extracted from evaluating a periodic function. The number of grid points will be referred to $N$ and it is *even*. The formulas can be redetermined for an *odd* $N$. The spacing of the grid points is: $h = 2\pi/N$, which is equivalent to:

$$\frac{\pi}{h} = \frac{N}{2} \tag{4.1}$$

The interval $[-\pi/h, \pi/h]$ is the range of wavenumbers distinguishable on the grid.

Let consider the Fourier transform on the $N$-point grid. The mesh spacing $h$ implies that wavenumbers differing by an integer multiple of $2\pi/h$ are indistinguishable on the grid, and thus it will be enough to confine our attention to $k \in [-\pi/h, \pi/h]$. The Fourier domain is discrete and bounded. Waves in physical space must be periodic over the interval $[0, 2\pi]$, and only waves $e^{ikx}$ with integer wavenumbers have the required period $2\pi$.

$$\begin{array}{lccc} \text{Physical space}: & \text{discrete} & \text{bounded}: & x \in \{h, 2\pi, \ldots, 2\pi - h, 2\pi\} \\ & \updownarrow & \updownarrow & \\ \text{Fourier space}: & \text{bounded} & \text{discrete}: & k \in \{-\frac{N}{2}+1, -\frac{N}{2}+2, \ldots, \frac{N}{2}\} \end{array}$$

The *Discrete Fourier Transform (DFT)*, for a function $v$ defined on $h, 2h, \ldots, 2\pi$ with value $v_j$ at $x_j$, is defined by

$$\hat{v_k} = h \sum_{j=1}^{N} e^{-ikx_j} v_j, \qquad k = -\frac{N}{2} + 1, \ldots, \frac{N}{2}, \tag{4.2}$$

The number $\hat{v_k}$ can be interpreted as the amplitude density of $v$ at wavenumber $k$. Conversely, the reconstruction of $v$ from $\hat{v_k}$ with the *inverse DFT* is:

$$v_j = \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2} e^{ikx_j} \hat{v_k}, \qquad j = 1, \ldots, N. \tag{4.3}$$

The variable $x$ is the *physical variable*, and $k$ is the *Fourier variable* or *wavenumber*. The $k$ and spatial index $j$ take only integer values. Equation (4.2) and (4.3) are inverses of one another for arbitrary vectors $(v_1, \ldots, v_N)^T \in \mathbb{C}^N$

A complication arises since the inverse transform (4.3) would give a term $e^{iNx/2}$ and derivative $(iN/2)e^{iNx/2}$. The term $e^{iNx/2}$ represents a real wave on the grid and its derivative should be zero at the grid points. However, a complex exponential is produced as the highest wavenumber is treated asymmetrically. After determining $\hat{v}$ and defining $v_{-\hat{N/2}} = v_{\hat{N/2}}$, the equation (4.3) is replaced by:

$$v_j = \frac{1}{2\pi} \sum_{k=-N/2}^{N/2} e^{ikx_j} \hat{v_k}, \qquad j = 1, \ldots, N, \tag{4.4}$$

where the prime indicates that the terms $k = N/2$ are multiplied by $\frac{1}{2}$. A band–limited interpolant is needed for the spectral differentiation and the inverse transform in (4.4) provides us one. The interpolant $p$ is a *trigonometric polynomial* of degree at most $N/2$ and is defined by

$$p(x) = \frac{1}{2\pi} \sum_{k=-N/2}^{N/2} e^{ikx} \hat{v_k}, \qquad x \in [0, 2\pi]. \tag{4.5}$$

The band–limited interpolant of delta function can be used to interpolate a grid function $v$. Then we can compute and expand $v$ as a linear combination of translated delta functions. The delta function is periodic:

$$\delta_j = \begin{cases} 1 & j \equiv 0 \pmod{N}, \\ 0 & j \not\equiv 0 \pmod{N}. \end{cases} \tag{4.6}$$

If we replace $\upsilon_j$ with $\delta_j$ in equation (4.2), we find that $\hat{\delta}_k = h, \forall k$. Thus equation (4.5) results in:

$$p(x) = \frac{h}{2\pi} \sum_{k=-N/2}^{N/2} e^{ikx} = \frac{h}{2\pi} \left( \frac{1}{2} \sum_{k=-N/2}^{N/2-1} e^{ikx} + \frac{1}{2} \sum_{k=-N/2+1}^{N/2} e^{ikx} \right) \tag{4.7a}$$

$$= \frac{h}{2\pi} \big( \frac{1}{2} \sum_{-N/2+1/2}^{N/2-1+1/2} e^{ikx+x/2} + \frac{1}{2} \sum_{-N/2+1-1/2}^{N/2-1/2} e^{ikx-x/2} \big) \tag{4.7b}$$

$$= \frac{h}{2\pi} \big( \frac{1}{2} \sum_{-N/2+1/2}^{N/2-1/2} e^{ikx} e^{x/2} + \frac{1}{2} \sum_{-N/2+1/2}^{N/2-1/2} e^{ikx} e^{-x/2} \big) \tag{4.7c}$$

$$= \frac{h}{2\pi} \big( \frac{1}{2} e^{ix/2} \sum_{-N/2+1/2}^{N/2-1/2} e^{ikx} + \frac{1}{2} e^{-ix/2} \sum_{N/2+1/2}^{N/2-1/2} e^{ikx} \tag{4.7d}$$

$$= \frac{h}{2\pi} \big( \sum_{-N/2+1/2}^{N/2-1/2} e^{ikx} \big( \frac{1}{2} (e^{ix/2} + e^{-ix/2}) \big) \tag{4.7e}$$

$$= \frac{h}{2\pi} \cos(x/2) \sum_{k=-N/2+1/2}^{N/2-1/2} e^{ikx} \tag{4.7f}$$

$$= \frac{h}{2\pi} \cos(x/2) \frac{e^{i(-N/2+1/2)x} - e^{i(N/2+1/2)x}}{1 - e^{ix}} \tag{4.7g}$$

$$= \frac{h}{2\pi} \cos(x/2) \frac{e^{-i(N/2)x} - e^{i(N/2)x}}{e^{-ix/2} - e^{ix/2}} \tag{4.7h}$$

$$= \frac{h}{2\pi} \cos(x/2) \frac{\sin(Nx/2)}{\sin(x/2)} \tag{4.7i}$$

If we apply the identity (4.1), the band–limited interpolant to $\delta$ is the *periodic sinc function* $S_N$ (Figure 4.2):

$$S_N(x) = \frac{\sin(\pi x/h)}{(2\pi/h)\tan(x/2)} \tag{4.8}$$

The $\delta_{j-m}$ can be interpolated as the delta function previously. Band–limited interpolation is a translation–invariant process in the sense that for any $m$, the band–limited interpolant of $\delta_{j-m}$ is $S_N(x - x_m)$. The grid function $\upsilon$ can be written as:

$$\upsilon_j = \sum_{m=1}^{N} \upsilon_m \delta_{j-m} \tag{4.9}$$

and the interpolant $p$:

$$p(x) = \sum_{m=1}^{N} \upsilon_m S_N(x - x_m) \tag{4.10}$$

We differentiate the interpolant $p$ of equation (4.10):

$$w_j = p'(x_j) = \sum_{m=1}^{N} \upsilon_m S'_N(x_j - x_m) \tag{4.11}$$

The differentiation of (4.10) generates:

$$S'_N(x_j) = \begin{cases} 0 & j \equiv 0 \pmod N, \\ \frac{1}{2}(-1)^j \cot(jh/2) & j \not\equiv 0 \pmod N. \end{cases} \tag{4.12}$$

The spectral differentiation matrix $N \times N$ is described by (4.12). The equation (4.11) is interpreted as a matrix equation and the vector $S'_N(x_j)$ is the column $m = 0$ of $D$, with the other columns obtained by shifting this column up or down appropriately. The $m$th column of a spectral differentiation matrix contains the values $p'(x_j)$, where $p(x)$ is the global interpolant through the discrete delta function supported at $x_m$:

$$D_N = \begin{pmatrix} 0 & & & -\frac{1}{2}\cot\frac{1h}{2} \\ -\frac{1}{2}\cot\frac{1h}{2} & \ddots & \ddots & \frac{1}{2}\cot\frac{2h}{2} \\ \frac{1}{2}\cot\frac{2h}{2} & \ddots & & -\frac{1}{2}\cot\frac{3h}{2} \\ -\frac{1}{2}\cot\frac{3h}{2} & & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{2}\cot\frac{1h}{2} \\ \frac{1}{2}\cot\frac{1h}{2} & & & 0 \end{pmatrix} \tag{4.13}$$

The matrix (4.13) is a toeplitz matrix. An example of the applied spectral differentiation to the function : $e^{\sin(x)}$ is shown in 4.1. The Spectral Derivative exhibits error of $10^{-13}$ compared to the conventional derivative.

The calculated toeplitz matrix is multiplied with the EGG signal. An example of the spectral differentiation on EGG signal is illustrated in Figure 4.3.

The calculation of higher spectral derivatives requires the differentiation of interpolant $p$ in equation (4.10) several times. For example, the second-order spectral differentiation is computed as:

$$S''_N(x_j) = \begin{cases} -\frac{\pi^2}{3h^2} - \frac{1}{6} & j \equiv 0 \pmod N, \\ -\frac{(-1)^j}{2\sin^2(jh/2)} & j \not\equiv 0 \pmod N, \end{cases} \tag{4.14}$$

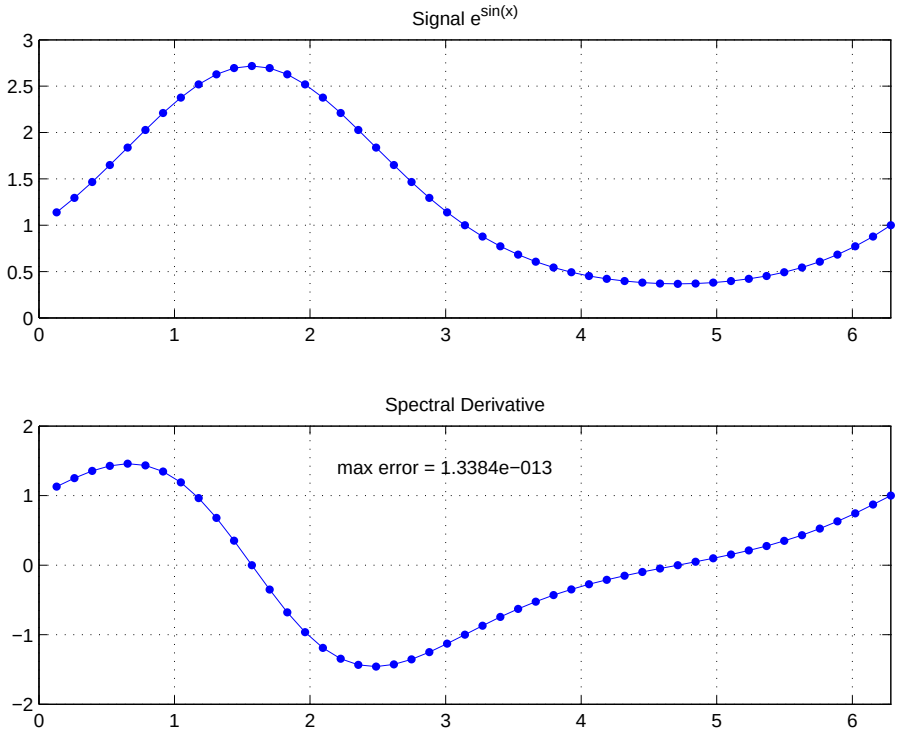



Figure 4.1: (a) Function $e^{\sin(x)}$ (b) The spectral derivative of $e^{\sin(x)}$

which can be written in the matrix form:

$$D_N^{(2)} v = \begin{pmatrix} \ddots & \vdots & \\ \ddots & -\frac{1}{2}\csc^2\left(\frac{2h}{2}\right) & \\ \ddots & \frac{1}{2}\csc^2\left(\frac{1h}{2}\right) & \\ & -\frac{\pi^2}{3h^2} - \frac{1}{6} & \\ & \frac{1}{2}\csc^2\left(\frac{1h}{2}\right) & \ddots \\ & -\frac{1}{2}\csc^2\left(\frac{2h}{2}\right) & \ddots \\ & \vdots & \ddots \end{pmatrix} v \tag{4.15}$$

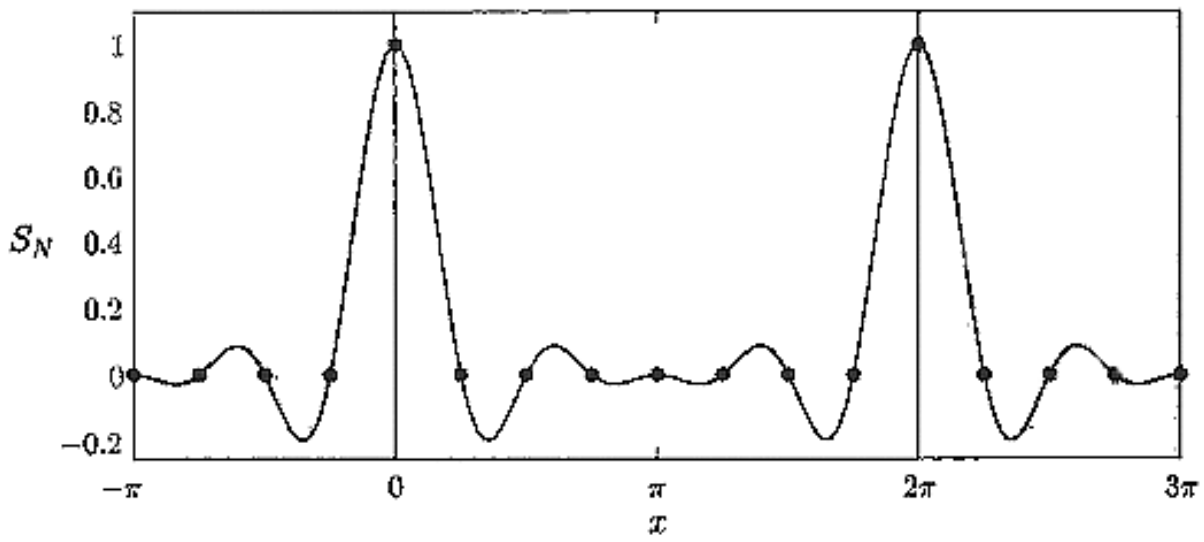



Figure 4.2: The periodic sinc function $S_N$, the band–limited interpolant of the periodic delta function $\delta$, plotted for $N = 8$.

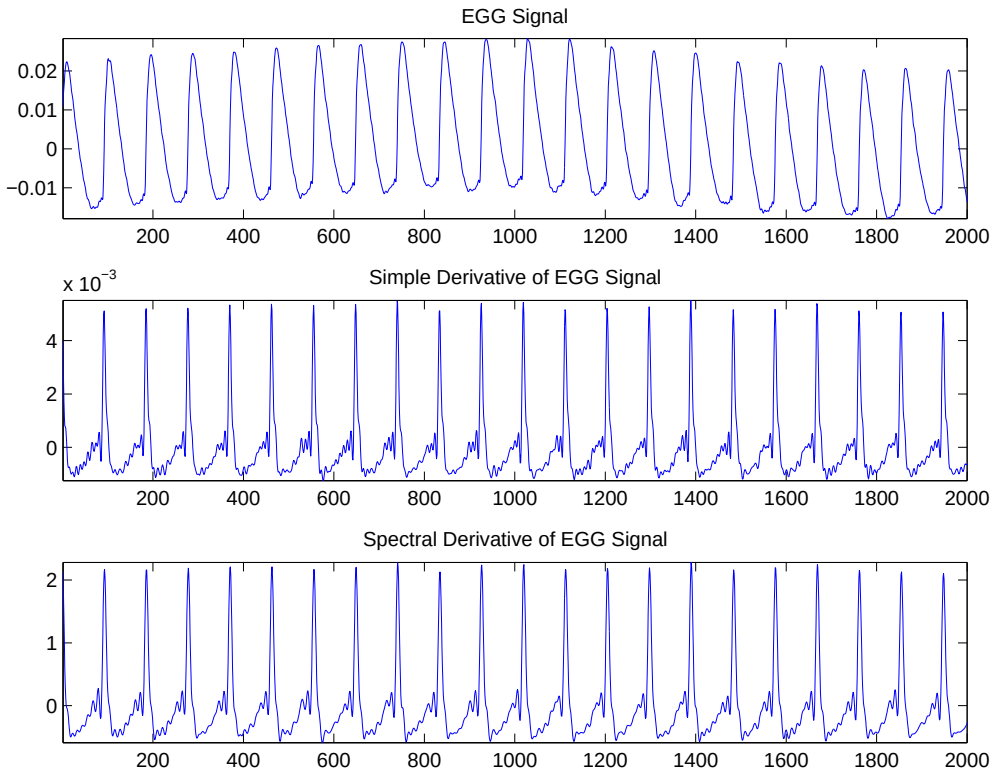



Figure 4.3: A sustained vowel /a/ recorded by a young female speaker. Illustration of the (a)EGG signal, (b) Simple Derivative of EGG signal, and (c)Spectral Derivative of EGG signal.

## 4.2 "Slope Filtering" Method

The "Slope Filtering" method has its origins in telecommunication and particularly in communications systems [55]. Nevertheless, this technique can be used effectively in several applications such as communications receiver carrier recovery, signal rate of change estimation, signal transition and transition-polarity detection.

The slope filtering method was evolved in communications systems, where the design of a carrier recovery process was essential. The communications' receiver and transmitter oscillators vary in frequency, though for short–time intervals they have nearly constant frequencies. The phase offset error function is derived from the instantaneous phase of the received signal and the phase of the ideal signal. The error function is approximately linear with respect to time. The phase offset error has to be estimated and minimized to demodulate the received signal. The linearity of the error function's noise directs a statistical method for its estimation.

The slope filtering method was initially developed from linear regression in the field of statistics for efficient computation of the phase error function. A set of $N$ ordered pairs $(x_0, y_0), (x_1, y_1), ..., (x_{N-1}, y_{N-1})$ define the data. There is no assumption of equal spacing for the $x_k$ values. If we fit, via a least squares method, a straight line through the data then the regression line is:

$$\hat{y} = \alpha + \beta\hat{x},$$

where

$$\alpha = \frac{\left(\sum_k y_k\right)\left(\sum_k {x_k}^2\right) - \left(\sum_k x_k\right)\left(\sum_k x_k y_k\right)}{N\left(\sum_k {x_k}^2\right) - \left(\sum_k x_k\right)^2} \tag{4.16}$$

$$\beta = \frac{N\left(\sum_k x_k y_k\right) - \left(\sum_k x_k\right)\left(\sum_k y_k\right)}{N\left(\sum_k {x_k}^2\right) - \left(\sum_k x_k\right)^2} \tag{4.17}$$

and $k = 0, 1, \ldots, N-1$.

The slope of the best–fit line to a $N$–length set of data is given by $\beta$. As it is shown in (4.17) the $\beta$ is dependent of $x_k, y_k$ and this relation renders restrictions for an efficient implementation in a signal processing application. Therefore the $y_k$ should be separated from $x_k$. First of all, the index $k$ of the summation over $y$ is replaced with a different index from that of $x$. The index $i$ is used for this purpose.

$$\beta = \frac{N\Big(\sum_i x_i y_i\Big) - \Big(\sum_k x_k\Big)\Big(\sum_i y_i\Big)}{N\Big(\sum_k {x_k}^2\Big) - \Big(\sum_k x_k\Big)^2} \tag{4.18}$$

and $k, i = 0, 1, \ldots, N-1$.

Then the $y_i$ samples are factored out:

$$\beta = \sum_i \left( \frac{N x_i - \sum_k x_k}{N\Big(\sum_k {x_k}^2\Big) - \Big(\sum_k x_k\Big)^2} \right) y_i \tag{4.19}$$

The equation (4.19) can be written as

$$\beta = \sum_i \beta_i y_i \tag{4.20}$$

where $\beta_i$ are defined as follows

$$\beta_i = \frac{N x_i - \sum_k x_k}{N\Big(\sum_k {x_k}^2\Big) - \Big(\sum_k x_k\Big)^2} \tag{4.21}$$

According to (4.19) if all $x_i$ samples are known a priori, then the $\beta_i$ coefficients may be precomputed. By inference, if the $N$–length $\beta_i$ coefficients are prestored for a set of $N$–length data, the computation of the slope of the data is computationally efficient.

Slope filtering is appropriate for applications in which continual rate of change estimations are essential. If the $x_k$ samples are equally spaced in an ascending order, then every sample can be rewritten as $x_k = x_0 + k$, where $k = 0, 1, 2, \ldots, N-1$. The $\beta_i$ are redefined from (4.21) :

$$\beta_i = \frac{N(x_0 + i) - \sum_k (x_0 + k)}{N\Big(\sum_k (x_0 + k)^2\Big) - \Big(\sum_k (x_0 + k)\Big)^2} \tag{4.22}$$

after algebraic simplification results in

$$\beta_i = \frac{12i - 6(N-1)}{N\left(N^2 - 1\right)} \tag{4.23}$$

The implemented (4.20) is designed for filter correlations. If the delay line structure is addressed for filter convolutions, the $\beta_i$ coefficients need to be reordered in a negative symmetry. The Figure 4.4 depicts a real–time slope filtering structure with a convolution delay line structure. The equation (4.23) describes the coefficients of an odd symmetric linear–phase finite impulse response (FIR) filter, which has a constant delay of $(N-1)/2$ samples. Furthermore, these coefficients can be precomputed and are independent of the data, since in (4.23) there is no $x$–input sample. Therefore $\beta_i$ are time-invariant coefficients and are determined by $N$. If the expected transition's duration is $M$ samples, the $N$ in the equation becomes $M$. Additionally, the longer width in samples than $M$ may increase SNR in the slope filtering output. The selection of the width should not exceed much. The duration of 20% to 50% greater than the transition width ($M$) is considered to be effective.

The slope filtering applied in signal transition detection is illustrated in Figure 4.5. A series of noisy input signals with transitions of 80 samples length are depicted with red color and the slope filtering output with black. In the first case, a short sample duration in slope filtering is used, $N = 21$, thus the output is considerably noisy. The second output is extracted from $N = 101$ samples, duration greater than the transition length of $\approx 26\%$. The result is significantly less noisy than the first. The last slope filtering output exhibits a reliable rate of change transmition signal, although the input signal contains greater noise than the previous inputs.

The ends of the data length are determinant for the regression method. On the contrary, the conventional differentiators accumulate the strongest weights near the middle. Therefore the time–domain slope filtering is noise tolerant. Moreover, another difference between the two approaches is the first–order polynomial applied in slope filtering method and the low–order trigonometric polynomial used in the conventional differentiators.

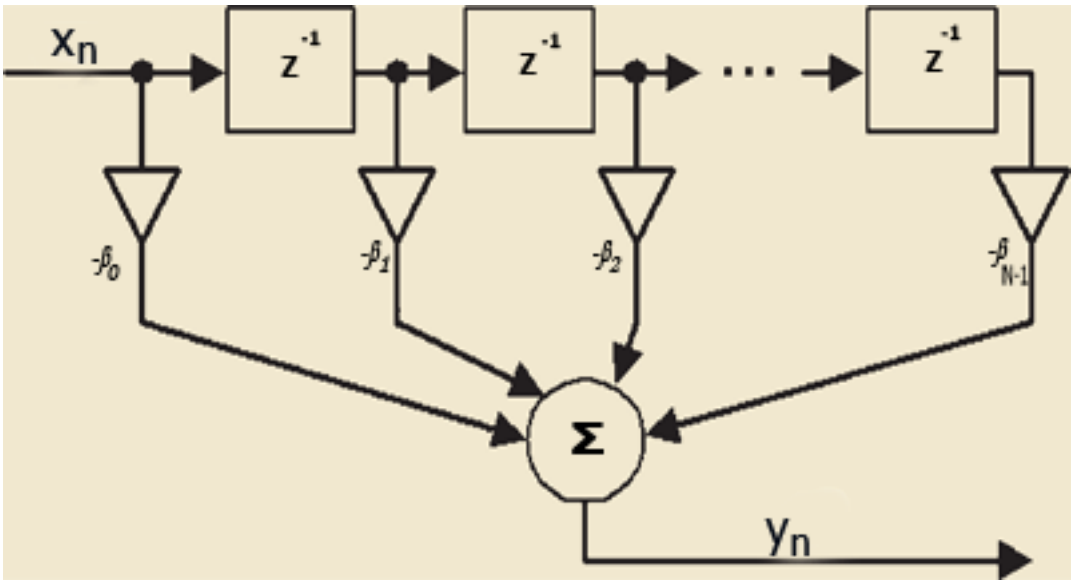



Figure 4.4: FIR real–time slope filtering structure

## 4.3 Data Processing

### 4.3.1 Database

Eleven male and five female young speakers participated in the experiments and were recorded in an anechoic chamber. Speech pressure waveform (Sp) and electrolaryngographic output waveform (Lx) were extracted with Laryngograph Ltd [32] and Speech Studio software. The microphone, which was used for the audio recording, was placed at a distance of $\approx$ 9 cm from the mouth. Particular care was taken with the positioning of the electrolaryngograph electrodes for a maximum amplitude Lx waveform during the procedure. The sampling rate was 16000 kHz. Each subject was asked to produce three separate recordings of each of the sustained vowels (/a/, /e/, /i/, /o/, /u/).

### 4.3.2 Align Speech with Egg Signal

The speech and EGG signals were time aligned to compensate for the larynx to microphone delay. There is a substantial physical distance between the microphone and laryngograph's electrodes placed at distant locations, thus there exists a time delay misalignment between the signals from different input channels. The computation of the distance between the larynx and the microphone and the air velocity define the time delay.

In addition, the speech and EGG signals are compared as illustrated in Figure 4.6 in order to confirm the alignment [24, 8]. If the glottal closure instant (GCI point 7) is defined as the point of maximal closing of vocal folds, then it is not possible to determine it from the speech signal itself, because there is no characteristic point in the speech signal which can be

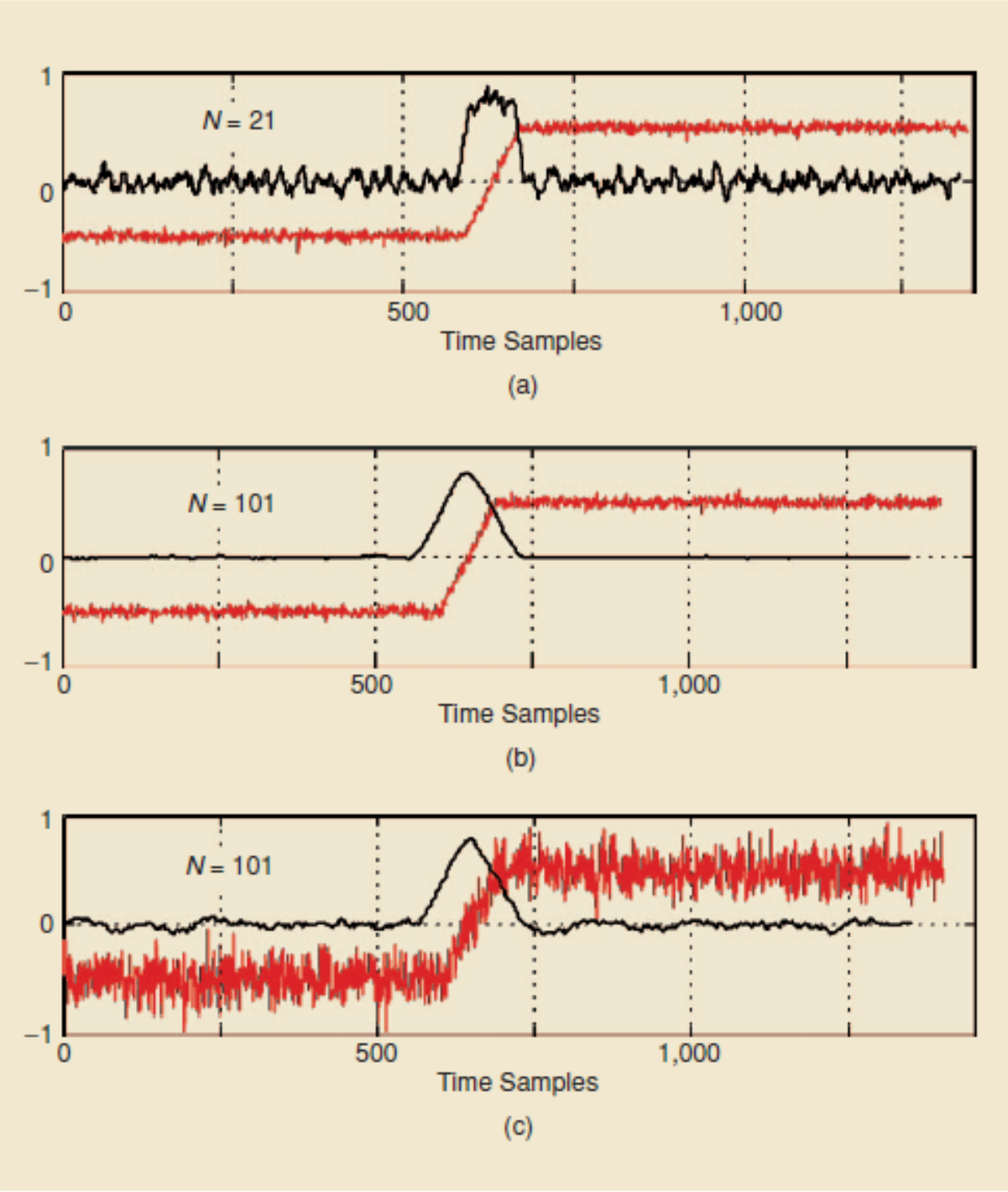



Figure 4.5: Slope filtering examples: (a) $N = 21$, (b) $N = 101$, and (c)$N = 101$, with a very noisy input signal

consistently declared as the point where the GCI occurs. The time instant at which the lower margins of the vocal folds are touching each other (point 6) can be uniquely and consistently determined on the basis of the speech signal itself.

The phases, delineated in Figure 4.6, are described below:

- **7-8:** Interval in which Vocal folds are closed.
- **8-1:** Vocal folds are separating from lower margins towards upper margins.
- **1-2:** Upper margins are opening.
- **2-3:** Upper margins are still opening. Changed slope is due to phase differences along the length of the vocal folds.
- **3-4-5:** Vocal folds are fully apart.
- **5-6:** Lower margins are approaching each other with a phase difference along the length of the vocal folds.
- **6:** Lower margins of vocal folds are touching each–other. This leads to sudden reduction of electrical resistance between the electrodes of the Laryngograph as well as to sudden change in the airflow through the glottis.
- **6-7:** Vocal folds are closing from lower margins towards upper margins.

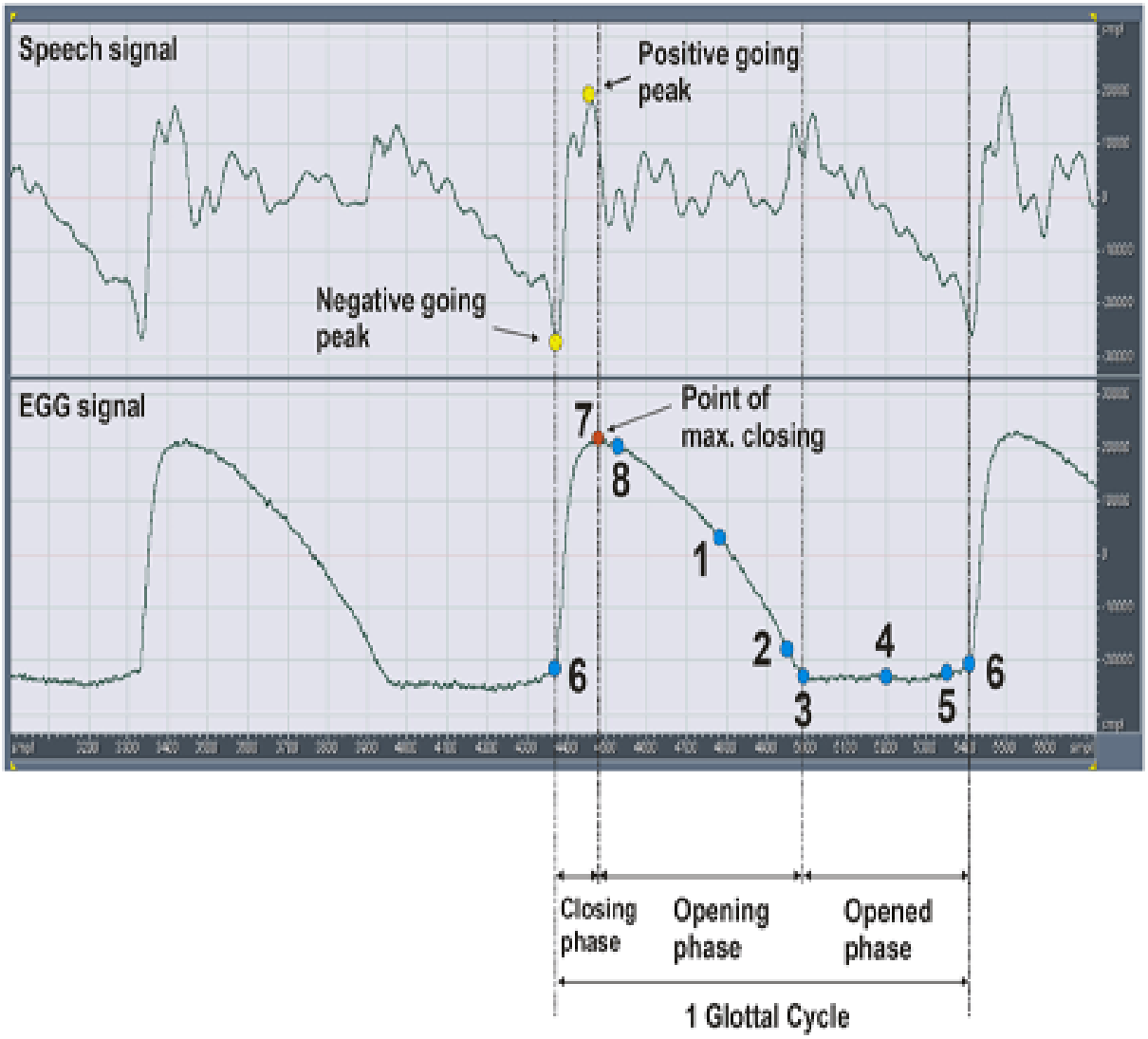



Figure 4.6: Comparison between Speech and EGG signal with glottal cycle markers

### 4.3.3 Selection of $N$ in slope Filter

The variable $N$ used in equation (4.23) depends on the transitional duration. In EGG signal the respective transition in samples has been observed to fluctuate from 10 to 15 samples. Therefore, the selection of $N$ would be close to this transition. The slope filtered EGG signal with different selections of $N$ is depicted in Figure 4.7. The simple derivative of EGG is illustrated in Figure 4.7a. Below the derivative, the EGG signal is filtered from the described FIR system for $N = 9, 11, 13$. The usage of $N = 9$ is not recommended since it does not cover the period of the transition and there are several negative peaks observed. Nevertheless, it is fundamental to appreciate the effect of $N$. The result for $N = 13$ is smoother than the previous used and the peaks are not well–distinguished. The $N = 11$ was preferred since the outcome approaches more effectively the conventional derivative. The positive peaks are discrete and the negative peaks are related to those in simple derivative. The delayed samples in the FIR system are $(N - 1)/2 = 5$. Moreover, the "slope filtering" can adjust to various transitions in EGG with different selections of $N$.

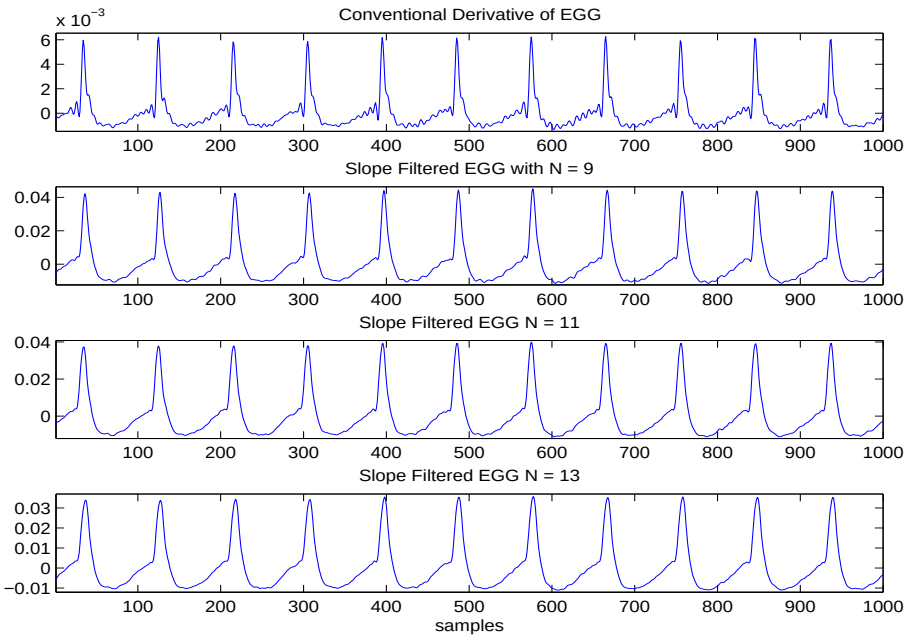



Figure 4.7: Illustration of the (a) simple derivative applied to the EGG signal, and "slope filtering" applied to EGG with different $N$ (b) $N = 9$, (c) $N = 11$, (d) $N = 13$

### 4.3.4 Thresholding Method

A differentiated signal displays two peaks for each glottal cycle, corresponded to the GCI and GOI. A method is required in order to extract the glottal Instants. The most widely implemented technique in DEGG signal is the usage of threshold. Large changes in amplitude of EGG can cause errors in dynamic threshold–based algorithms. If the threshold gain is set

too high, sometimes missed peaks occur and if it is set too low, erroneous Instants from noise are estimated. The pick of the appropriate threshold is determinant as it concerns the GCI and the relevant positive peak in differentiated signal. In addition, the right minimum threshold should be set for the GOI and the corresponded negative peak. Particular interest attracts the onset and offset of voiced speech, since the majority of approaches lack in the determination of the glottal Instants in these areas. However, the proposed method is robust to these misidentifications. The input signal is divided into segments to estimate efficiently the local maximum and minimum. Thus, maximal and minimal thresholds are assigned for each segment. This process is essential as any differentiated signal consists of many variations in amplitude. A total threshold (e.g. maximal) would cause the miss of lower peaks in an interval, in which the signal shows weaker peaks.

Alternative thresholds have been applied and tested compared to the Speech Filing System, which follows a thresholding technique. In this thesis, the threshold was set at $40\%$ for GCI and GOI location. This threshold supported the determination of both peaks in double–peak case (3.2.1) as illustrated in Figure 4.8.

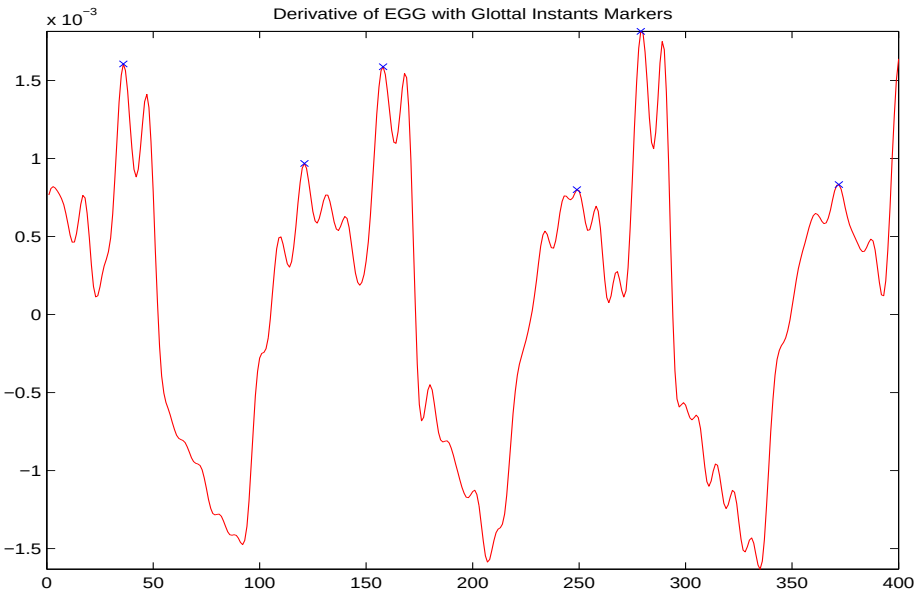



Figure 4.8: Detection of double closure peaks (marked with '*')

A thresholding approach combined with the spectral derivative exhibits similar efficiency to those with the simple derivative. The filtered signal from "slope" method is more robust to noise and a thresholding approach generates more reliable estimations than applied to the other differentiation ways.

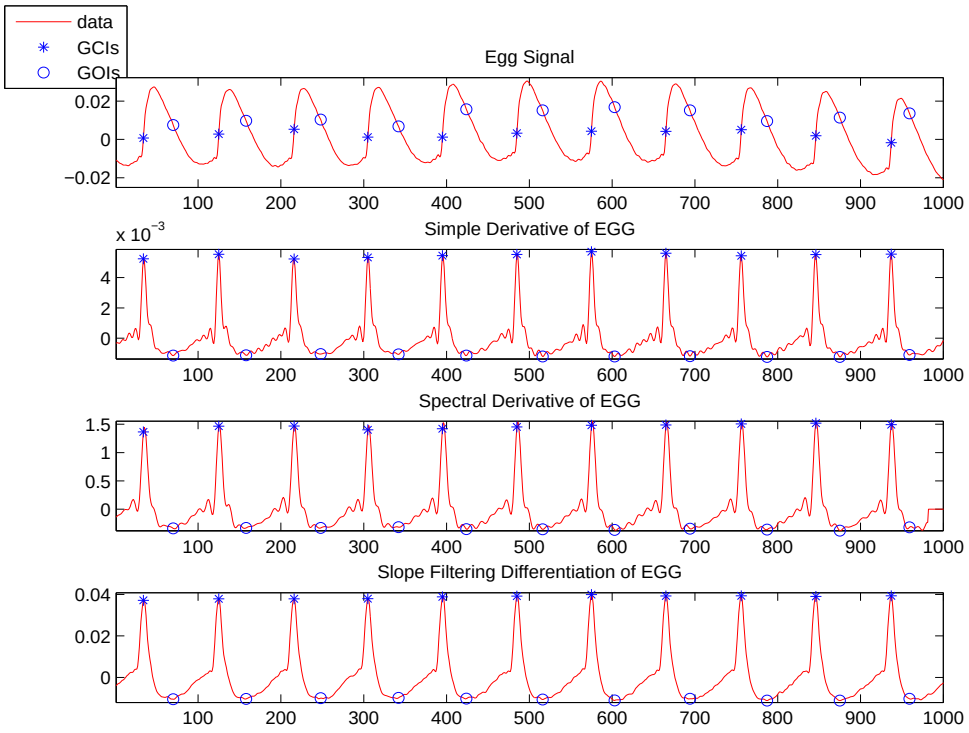



Figure 4.9: Illustration of the (a) EGG signal, (b) Simple Derivative of EGG and (c) Spectral Derivative of EGG signal, and (d) Slope Filtered EGG for a young female speaker and the phoneme /a/ (GCIs-⋆ markers, GOIs- ∘ markers)

## 4.4 Evaluation

The EGG signal is aligned to the speech waveform as it was described above and it is low–passed at $3000Hz$. Then, it is differentiated with spectral approximation or "slope filtering" and a "Thresholding" technique is applied to determine the glottal Instants. The evaluation of the presented methods involves the computation of the simple derivative and the "thresholding" extraction of the glottal information. The generated Instants from the conventional derivative are the referenced ones for the experiments that follow. The assessment of the performance of the methods is based on the measures defined in [39] and pictured in Figure 4.10. These measures include:

- **Identification Rate (IDR):** the percentage of larynx cycles for which exactly on glottal instant (GCI/GOI) is detected
- **Miss Rate (MR):** the percentage of larynx cycles for which no GCI/GOI is detected
- **False Alarm Rate (FAR):** the percentage of larynx cycles for which more than one GCI/GOI is detected
- **Identification Error:** the timing error between the reference GCI/GOI and the detected GCIs/GOIs in the cycles for which exactly one Instant has been detected (-$\zeta = (t_d(i) - t_{ref}(i))$, $t_d(i)$ is the detected Instant in larynx cycle $i$ and $t_ref(i)$ is the reference Instant in larynx cycle $i$ )

- **Identification Accuracy (IDA):** the standard deviation of the error (of $\zeta$)
- **Mean Error:** the mean value of error $\zeta$ is computed for correction
- **Accuracy:** the rate of detections for which the timing error is smaller than the defined bound ($\pm 0.25$ms)

Additionally, the F-Measure was used, which can be interpreted as a weighted average of the precision and recall. F-Measure ranges in the interval [0,1] and reaches its best value at 1 and worst score at 0.:

$$F = \frac{2 \times precision \times recall}{(precision + recall)}$$

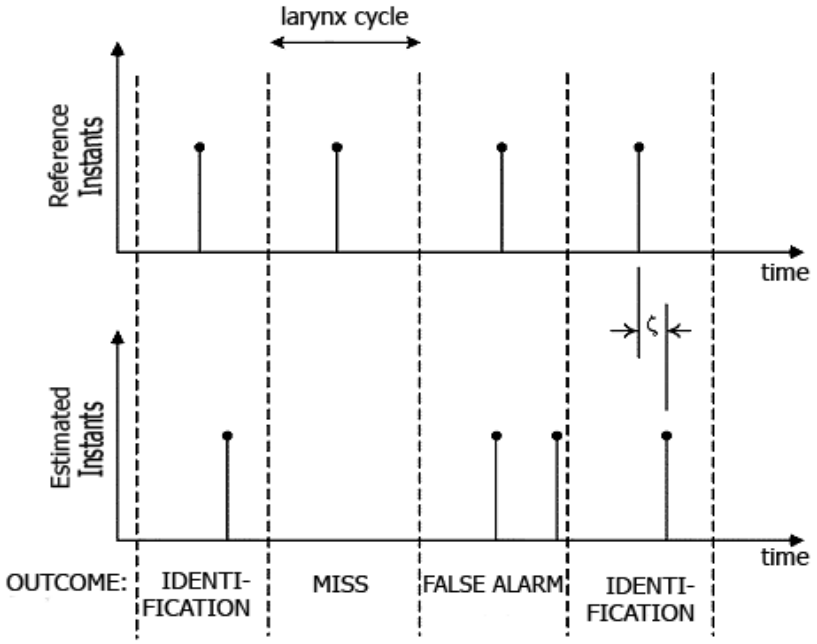



Figure 4.10: Characterization of Glottal Instants with examples of each possible case from their estimation for each larynx cycle.

The proposed methods and the simple derivative are compared based on their noise robustness. A white Gaussian noise is added to the EGG signal in different levels. The Signal–to–Noise Ratio (SNR) varies from 25dB to 40dB. The glottal Instants that are detected through a "thresholding" technique in DEGG signal (simple derivative), before noise added to EGG, are used as the referenced Instants. The distorted EGG signal is used as input to the simple derivative, to the spectral differentiation and to "slope filtering" method. The results are shown in the following tables. Tables 4.1 and 4.3 indicate the F-Measure for all phonemes aggregated for all men and for all women, respectively, in the database. The F-Measure of each method increases, as the SNR increases as well. The tables demonstrate that the slope filtering method outperforms the compared algorithms. More specifically, tables 4.6 and 4.5 show the rates of identification, miss, false alarm and identification accuracy for female and

male speakers in EGG signals with additive noise of SNR = 40dB. It is observed that the slope filtering process exhibits better efficiency, despite the fact that the same thresholding algorithm has been applied to all comparative algorithms. It is noticeable that except for the high Identification Rate of the slope filtering algorithm, the Miss Rate and False Alarm Rate are very low.

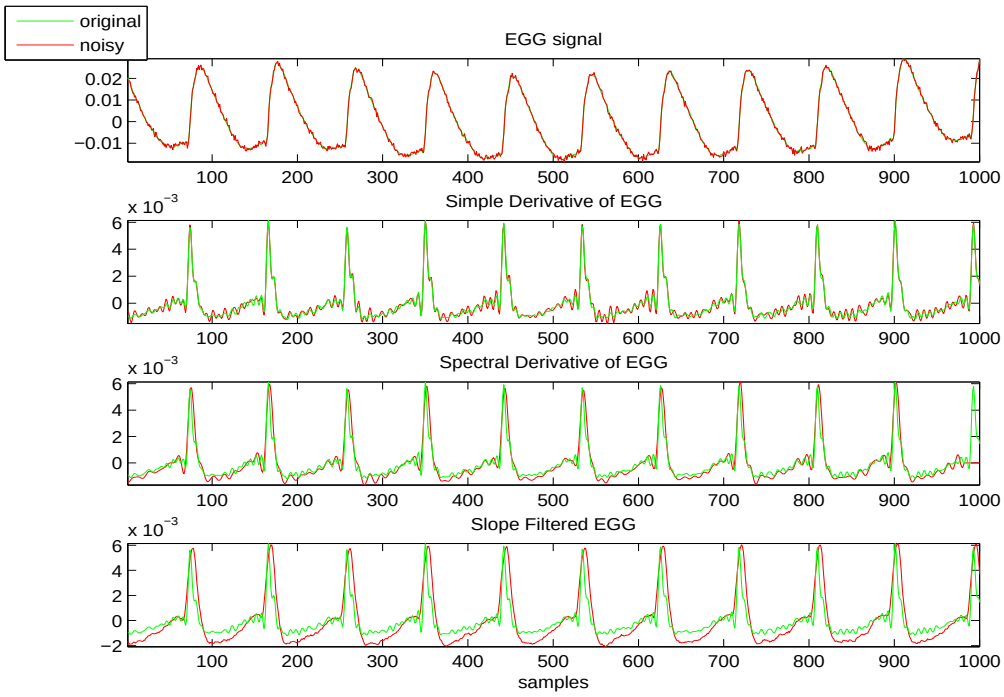



Figure 4.11: Illustration of original EGG signal and its simple derivative (green color) and the distorted (SNR = 25dB): (a)EGG signal, (b) Simple Derivative, (c) Spectral Derivative, (d) Slope Filtering Differentiation

Table 4.1: Comparative Results in terms of F-Measure of Male speakers for all phonemes

| **F-Measure, Male** | | | | | |
|---|---|---|---|---|---|
| **Event** | **Method** | **SNR=25dB** | **30dB** | **40dB** | **100dB** |
| | simple | 0.9974 | 0.9974 | 0.9977 | 1 |
| GCI | spectral | 0.8749 | 0.8950 | 0.9190 | 0.9190 |
| | slope | 0.9982 | 0.9984 | 0.9984 | 0.9985 |
| | simple | 0.9928 | 0.9932 | 0.9970 | 1 |
| GOI | spectral | 0.8663 | 0.8780 | 0.9083 | 0.9143 |
| | slope | 0.9966 | 0.9971 | 0.9971 | 0.9971 |

Table 4.2: F-Measure difference in Male speakers from SNR = 100dB to 25dB

| **Event** | **Method** | **Difference in F-Measure** |
|---|---|---|
| | simple | 0.0026 |
| GCI | spectral | 0.0441 |
| | slope | 0.0003 |
| | simple | 0.0072 |
| GOI | spectral | 0.0480 |
| | slope | 0.0005 |

Table 4.3: Comparative Results in terms of F-Measure of Female speakers for all phonemes

| **F-Measure, Female** | | | | | |
|---|---|---|---|---|---|
| **Event** | **Method** | **SNR=25dB** | **30dB** | **40dB** | **100dB** |
| | simple | 0.9877 | 0.9948 | 0.9950 | 1 |
| GCI | spectral | 0.9370 | 0.9372 | 0.9508 | 0.9575 |
| | slope | 0.9887 | 0.9951 | 0.9951 | 0.9951 |
| | simple | 0.9877 | 0.9948 | 0.9950 | 1 |
| GOI | spectral | 0.9602 | 0.9667 | 0.9732 | 0.9766 |
| | slope | 0.9950 | 0.9951 | 0.9951 | 0.9951 |

Table 4.4: F-Measure difference in Female speakers from SNR = 100dB to 25dB

| **Event** | **Method** | **Difference in F-Measure** |
|---|---|---|
| | simple | 0.0123 |
| GCI | spectral | 0.0205 |
| | slope | 0.0064 |
| | simple | 0.123 |
| GOI | spectral | 0.164 |
| | slope | 0.0001 |

Table 4.5: Comparative results in terms of IDR (%), MR (%), FAR (%), IDA (ms) and Accuracy to $\pm 0.25ms(\%)$ for male speakers

| **Male, SNR = 40dB** | | | | | | |
|---|---|---|---|---|---|---|
| **Event** | **Method** | **IDR** | **MR** | **FAR** | **IDA** | **Accuracy** |
| GCI | simple | 99.78 | 0.15 | 0.06 | 0.09 | 99.67 |
| | spectral | 70.83 | 28.75 | 0.41 | 0.28 | 69.49 |
| | slope | 99.27 | 0.3 | 0.41 | 0.62 | 90.51 |
| GOI | simple | 99.38 | 0.6 | 0.007 | 0.22 | 96.4 |
| | spectral | 78.38 | 21.56 | 0.04 | 0.55 | 73.88 |
| | slope | 99.42 | 0.57 | 0.0001 | 0.4 | 96.33 |

Table 4.6: Comparative results in terms of IDR (%), MR (%), FAR (%), IDA (ms) and Accuracy to $\pm 0.25ms(\%)$ for female speakers

| **Female, SNR = 40dB** | | | | | | |
|---|---|---|---|---|---|---|
| **Event** | **Method** | **IDR** | **MR** | **FAR** | **IDA** | **Accuracy** |
| GCI | simple | 98.99 | 1.00 | 0.002 | 0.07 | 98.96 |
| | spectral | 90.62 | 9.37 | 0.007 | 0.03 | 90.52 |
| | slope | 99.01 | 0.97 | 0.002 | 0.08 | 97.19 |
| GOI | simple | 98.98 | 1.01 | 0.0007 | 0.26 | 96.27 |
| | spectral | 93.54 | 6.44 | 0.01 | 0.49 | 83.42 |
| | slope | 99.01 | 0.98 | 0.001 | 0.39 | 90.44 |

## 4.5 Conclusions

In this chapter we proposed two methods to perform the gradient of electroglotographic signal;spectral and slope filtering. Experiments conducted with the described database. Comparison of the methods with equivalent and widely adopted approach showed that the spectral differentiation demonstrates similar behavior to the simple derivative. In terms of visibility, both of differentiations lack in exposing a clear, strong negative peak in common cases. The generated results from both methods range in close ratios. On the contrary, the slope filtering technique applied to the EGG signal offers great visibility of the major positive and negative peaks used for the detection of GCIs and GOIs. This approach proved to be more robust in revealing the GCIs and the GOIs, even in cases of noisy EGG recordings (SNR = 25dB, 30dB, 40dB). The technique relies on equations derived from linear regression and offers an estimation of the gradient of the EGG signal. In literature, the detection of GOIs has not attracted much interest as the location of GCIs. The identification of GOIs from the simple derivative using a thresholding approach meet difficulties in common cases, due to the presence of many negative peaks with similar amplitude. In slope filtered EGG signal the peaks are well distinguished and the negative peak can be uniquely specified. Therefore, the thresholding technique brings better results combined with slope filtering differentiator. Furthermore, the fact that the negative peak is more discrete than in simple derivative provides a more reliable indicator of GOI.

# Chapter 5

# High Resolution Speech Analysis

This chapter describes the relation between glottal phases and speech signal. The adaptive Quasi–Harmonic Model is applied to the speech signal, producing the AM and FM waveforms. The methodology is analyzed in the first part of the current chapter. Afterwards, the generated waveforms are compared to GCI and GOI that are derived from EGG signal through slope filtering process and the relative observations are discussed in the last part of the chapter. The Amplitude and Frequency Modulations of First and Second Formant are studied during glottal open and closed phases. The glottal activity is related to the AM–FM components of the speech signal.

## 5.1 Quasi–Harmonic Model of AM–FM Decomposition

In [48] (chapter 4) the quasi–harmonic component of a speech signal, within an analysis window, is modeled as

$$s(t) = \left( \sum_{k=-K}^{K} (a_k + t b_k) e^{j2\pi k f_0 t} \right) w(t) \tag{5.1}$$

where $f_0$ is the fundamental frequency of the harmonic signal
$K$ specifies the order of the model, i.e., the number of harmonics
$a_k$ are the complex amplitudes
$b_k$ are the complex slopes
$w(t)$ denotes the analysis window which is typically a Hamming window and zero outside a symmetric interval $[-T, T]$.]

Thus, $t = 0$ will always denote the center of the analysis window. In real signals such as speech, audio, etc., $a_{-k} = a_k^*$ and $b_{-k} = b_k^*$ where $^*$ is the conjugate operator. The model in (5.1) can also be written for non-harmonically related frequency components. Therefore, a

more general model can be expressed as

$$s(t) = \left( \sum_{k=-K}^{K} (a_k + tb_k) e^{j2\pi f_k t} \right) w(t) \tag{5.2}$$

where $f_k$ will be referred to as initial estimates of the frequencies.

Assuming that the speech signal $x(t)$ is defined on $[-T, T]$, the estimation of the model parameters $\{f_0, K, a_k, b_k\}$ is performed into two steps. At first, the fundamental frequency, $f_0$ and the number of harmonic components, $K$, are estimated using spectral and autocorrelation information as described in [48]. Then, the computation of $\{a_k, b_k\}_{k=-K}^{K}$ is performed by minimizing a mean squared error which leads to a simple least squares solution.

From (5.2), the instantaneous amplitude of each component is a time-varying function given by

$$\begin{aligned} M_k(t) &= |a_k + tb_k| \\ &= \sqrt{(a_k^R + tb_k^R)^2 + (a_k^I + tb_k^I)^2} \end{aligned} \tag{5.3}$$

where $x^R$ and $x^I$ denote the real and the imaginary parts of $x$, respectively.

Since both the amplitudes $a_k$ and the slopes $b_k$ are complex variables, the instantaneous frequency of each component is not a constant function over time but varies according to

$$\begin{aligned} F_k(t) &= \frac{1}{2\pi} \frac{d\Phi_k(t)}{dt} \\ &= f_k + \frac{1}{2\pi} \frac{a_k^R b_k^I - a_k^I b_k^R}{M_k^2(t)} \end{aligned} \tag{5.4}$$

while the instantaneous phase is given by

$$\begin{aligned} \Phi_k(t) &= 2\pi f_k t + \angle(a_k + tb_k) \\ &= 2\pi f_k t + atan \frac{a_k^I + tb_k^I}{a_k^R + tb_k^R} \end{aligned} \tag{5.5}$$

The Fourier transform of $s(t)$ in (5.2) is described by

$$S(f) = \sum_{k=1}^{K} \left( a_k W(f - f_k) + \frac{jb_k}{2\pi} W'(f - f_k) \right) \tag{5.6}$$

where $W(f)$ is the Fourier transform of the analysis window, $w(t)$, and $W'(f)$ is the derivative of $W(f)$ over $f$. For simplicity, the $k$th component of $S(f)$ is considered for the further calculations.

$$S_k(f) = a_k W(f - f_k) + \frac{jb_k}{2\pi} W'(f - f_k) \tag{5.7}$$

The projection of $b_k$ onto $a_k$ is defined by

$$b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k \tag{5.8}$$

where $ja_k$ denotes the perpendicular (vector) to $a_k$, while $\rho_{1,k}$ and $\rho_{2,k}$ are computed as

$$\rho_{1,k} = \frac{a_k^R b_k^R + a_k^I b_k^I}{|a_k|^2} \tag{5.9}$$

and

$$\rho_{2,k} = \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2} \tag{5.10}$$

Thus, the $k$th component of $S_k(f)$ can be written as

$$\begin{aligned} S_k(f) = \; & a_k W(f - f_k) - \frac{a_k \rho_{2,k}}{2\pi} W'(f - f_k) \\ & + \frac{j a_k \rho_{1,k}}{2\pi} W'(f - f_k) \end{aligned} \tag{5.11}$$

The Taylor series expansion of $W(f - f_k - \frac{\rho_{2,k}}{2\pi})$ is defined as

$$\begin{aligned} & W(f - f_k - \frac{\rho_{2,k}}{2\pi}) = \\ & = W(f - f_k) - \frac{\rho_{2,k}}{2\pi} W'(f - f_k) + O(\rho_{2,k}^2 W''(f - f_k)) \end{aligned} \tag{5.12}$$

For a rectangular window it holds that $W''(f) \propto T^3$ where $T$ is the duration of the analysis window, $w(t)$. Since the duration of the analysis window determines its bandwidth, it turns out that the larger the bandwidth the smaller the value of the term $W''(f)$ at $f_k$. Thus, assuming short analysis windows and low values for $\rho_{2,k}$, (5.12) is approximated as

$$W(f - f_k - \frac{\rho_{2,k}}{2\pi}) \approx W(f - f_k) - \frac{\rho_{2,k}}{2\pi} W'(f - f_k) \tag{5.13}$$

Consequently, from (5.11) and (5.13) it follows that

$$S_k(f) \approx a_k \left[ W(f - f_k - \frac{\rho_{2,k}}{2\pi}) + j \frac{\rho_{1,k}}{2\pi} W'(f - f_k) \right] \tag{5.14}$$

which is written in the time domain as

$$s_k(t) \approx a_k \left[ e^{j(2\pi f_k + \rho_{2,k})t} + \rho_{1,k} t e^{j2\pi f_k t} \right] w(t) \tag{5.15}$$

From (5.15), it is clear that $\frac{\rho_{2,k}}{2\pi}$ accounts for the mismatch between the frequency of the $k$th component and the initial estimate of the frequency, $f_k$, while $\rho_{1,k}$ accounts for the normalized amplitude slope of the $k$th component. Another way to see this relationship, is to associate the time domain and the frequency domain properties of QHM. From (5.4) and (5.10) it follows that

$$\frac{\rho_{2,k}}{2\pi} = F_k(0) - f_k \tag{5.16}$$

Therefore, $\frac{\rho_{2,k}}{2\pi}$ accounts for a frequency deviation between the initially estimated frequency, $f_k$, and the value of the instantaneous frequency at the center of the analysis window ($t = 0$). Similarly, for $\rho_{1,k}$, we have

$$\rho_{1,k} = \frac{\frac{dM_k(t)}{dt}|_{t=0}}{M_k(0)} \tag{5.17}$$

which shows that $\rho_{1,k}$ provides the normalized slope of the amplitude for the $k$th component, considering the instantaneous amplitude at the center of the analysis window.

## 5.2 Adaptive AM–FM Decomposition

Initialization of aQHM is provided by QHM. The updated frequencies, corresponded amplitudes and phases of time are denoted by $\hat{f}_k(t_l)$, $\hat{A}_k(t_l)$, and $\hat{\phi}_k(t_l)$, at time instant $t_l$ (center of analysis window), with $l = 1, \ldots, L$ where $L$ is the number of frames. These parameters are estimated using QHM as follows:

$$\hat{f}_k(t_l) = F_k(0) = f_k(t_l) + \frac{\rho_{2,k}}{2\pi} \tag{5.18a}$$

$$\hat{A}_k(t_l) = M_k(0) = |a_k| \tag{5.18b}$$

$$\hat{\phi}_k(t_l) = \Phi_k(0) = \angle a_k \tag{5.18c}$$

In case the distance between the consecutive analysis time instants correspond to one sample then, QHM provides an estimation of the instantaneous amplitude, $\hat{A}_k(t)$ and instantaneous phase $\hat{\phi}_k(t)$. Then, in aQHM the signal model is given as:

$$s(t) = \left( \sum_{k=-K}^{K} (a_k + t b_k) e^{j(\hat{\phi}_k(t-t_l) - \hat{\phi}_k(t_l))} \right) w(t) \tag{5.19}$$

with $|t| \leq T$ where $2T$ denotes the duration of the analysis window.

At the last step of aQHM, the signal can be finally approximated as the sum of its AM-FM components:

$$\hat{x}(t) = \sum_{k=-K}^{K} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{5.20}$$

Based on the definition of phase, the instantaneous phase for the $k$th component can be computed as the integral of the computed instantaneous frequency. For instance, between two consecutive analysis time instants $t_{l-1}$ and $t_l$, the instantaneous phase can be computed as:

$$\check{\phi}_k(t) = \hat{\phi}_k(t_{l-1}) + \int_{t_{l-1}}^{t} 2\pi \hat{f}_k(u) du \tag{5.21}$$

The frame boundary conditions at $t_l$ are not taken into account, which means that there is no guarantee that $\check{\phi}_k(t_l) = \hat{\phi}_k(t_l) + 2\pi M$, where $M$ is the closet integer to $|\hat{\phi}_k(t_l) - \check{\phi}_k(t_l)|/(2\pi)$. The modification of (5.21) in order to guarantee phase continuation over frame boundaries results in:

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_{l-1}) + \int_{t_{l-1}}^{t} 2\pi \hat{f}_k(u) + a \sin\left( \frac{\pi(u - t_{l-1})}{t_l - t_{l-1}} \right) du \tag{5.22}$$

The derivative of the instantaneous phase over time in both formulas provide the instantaneous frequency computed from $t_{l-1}$ to $t_l$. In (5.22), the continuation of instantaneous frequency

at the frame boundaries is also guaranteed by the use of the sine function. Using (5.22) the instantaneous phase at $t_l$ will be equal to $\hat{\phi}_k(t_l) + 2\pi M$ if $a$ is selected to be:

$$a = \frac{\pi(\hat{\phi}_k(t_l) + 2\pi M - \check{\phi}_k(t_l))}{2(t_l - t_{l-1})} \tag{5.23}$$

where $M$ is computed as before.

A pseudocode of the aQHM algorithm is presented below. **Adaptive AM-FM decomposition alg. (aQHM)**

1. Initialization (*QHM*):
   Provide initial estimate $f_k^0(t_1)$
   For $l = 1, 2, ..., L$
   (a) Compute $a_k$, $b_k$ using $f_k^0(t_l)$ as initial frequency estimates in (5.2)
   (b) Update $\hat{f}_k^0(t_l)$ using (5.18a) and
   (c) Compute $\hat{A}_k^0(t_l)$ and $\hat{\phi}_k^0(t_l)$ using (5.18b) and (5.18c), respectively
   (d) $f_k^0(t_{l+1}) = \hat{f}_k^0(t_l)$

   end
   Interpolate $\hat{f}_k^0(t)$, $\hat{A}_k^0(t)$, $\hat{\phi}_k^0(t)$ as described

2. Adaptations:
   For $i = 1, 2, ...$
   For $l = 1, 2, ..., L$
   a) Compute $a_k$, $b_k$ using $\hat{\phi}_k^{i-1}(t)$ and (5.19)
   b) Update $\hat{f}_k^i(t_l)$ using (5.18a) and (5.10)
   c) Compute $\hat{A}_k^i(t_l)$ and $\hat{\phi}_k^i(t_l)$ using (5.18b) and (5.18c), respectively
   end
   Interpolate $\hat{f}_k^i(t)$, $\hat{A}_k^i(t)$, $\hat{\phi}_k^i(t)$ as described
   end

## 5.3 Correlation between the Glottal Instants and the AM–FM Components

The adaptive QHM algorithm projects the input signal in a space generated by time varying non–parametric sinusoidal basis functions. The non-parametric basis functions are updated iteratively, minimizing the mean squared error at each iteration.

The peaks that are observed in the spectrum envelope, formants, for a certain vowel have a similar behavior. For instance, the first and second formant in vowels /a/, /u/ are closer than in the vowel /i/ for all speakers. A short–time Hamming window has been applied in aQHM algorithm depended on each recorded vowel. The algorithm is initialized with a formant's frequency. Two formants have been mainly studied; the first and second. The frequency of each formant is estimated through root–solving of LPC coefficients.

The descripted database in 4.3.1 was used in the experiments with aQHM method. The EGG signal is time–aligned with the speech signal and glottal Instants are derived from EGG signal. The glottal phases are depicted with the AM–FM components to derive the desired information of their variation within a glottal cycle.

The following Figures depict the aQHM algorithm applied in a young Female's speech signals. Each Figure shows the recorded speech signal, the slope filtered EGG signal, the EGG signal and the AM, FM components derived from aQHM method. The GCIs, GOIs are extracted with slope filtering processing of time–aligned EGG signal. The closed and open phases are illustrated in all signals with red and green color respectively. Figure 5.1 refers to the recorded sustained vowel /i/ and indicates the Amplitude, Frequency Modulations of First Formant ($F1$). Figure 5.2 includes the extracted components from the First Formant of sustained vowel /o/ and Figure 5.3 contains the information of the second Formant of vowel /o/. Figure 5.4 comprises the modulations of the First Fromant of sustained vowel /u/.

The Figure 5.5 depicts the same signals and modulations as in previous figures for a young male speaker and sustained vowel /u/.

The closed phase of glottis is noticed in the Amplitude Component of speech signal near its maximal values, as it is shown in Figure 5.6. The open phase of glottis is related to the minimum values of Amplitude. Concluding from the observations, when the glottis closes, the amplitude is increased and when the glottis opens, the amplitude drops. On the contrary, the Frequency Modulation exposes a different activity. The formant frequency is not related with glottal phases in the same way for all speech signals. However, a general pattern is followed. It is observed that the formant frequency rises at the onset of the glottal open phase and decreases near the end of the phase. It is noticeable that there is a similar variation for all the recorded sustained vowels.

The second Formant indicates similar modulations to that of the first formant. The glottal phases are related to the AM–FM components of second formant in the same way as in the first.

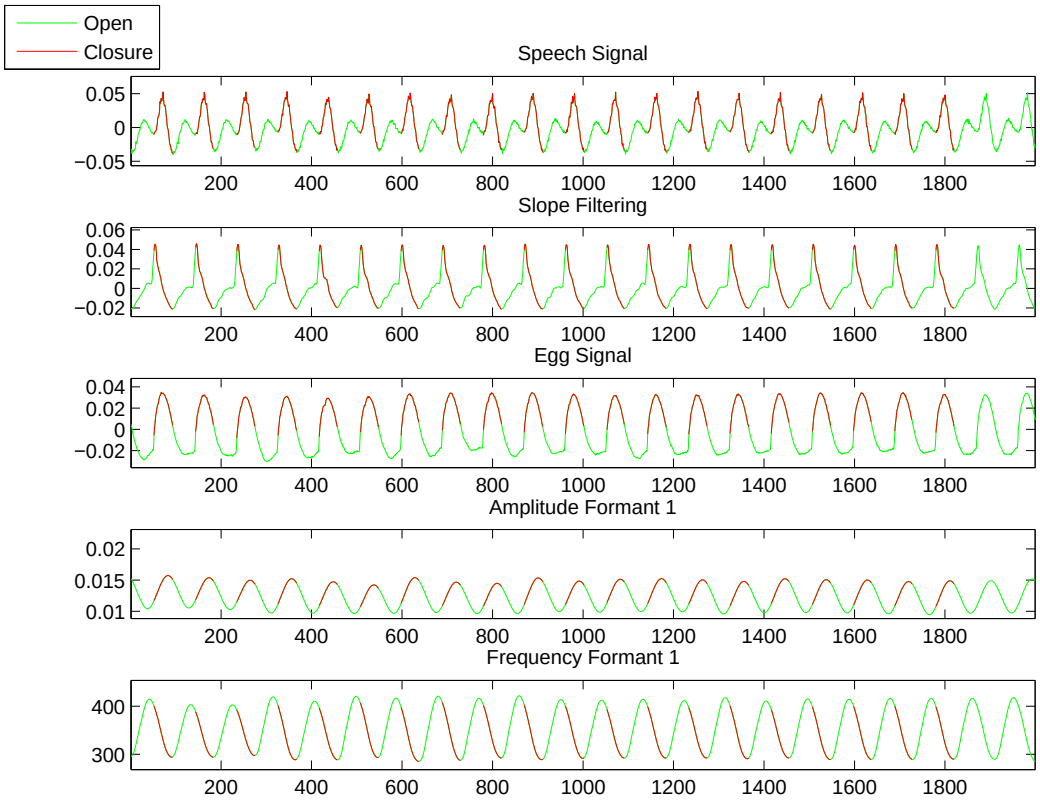



Figure 5.1: Recording of sustained vowel /i/ from Female Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 1st Formant, (e) Frequency Component of 1st Formant

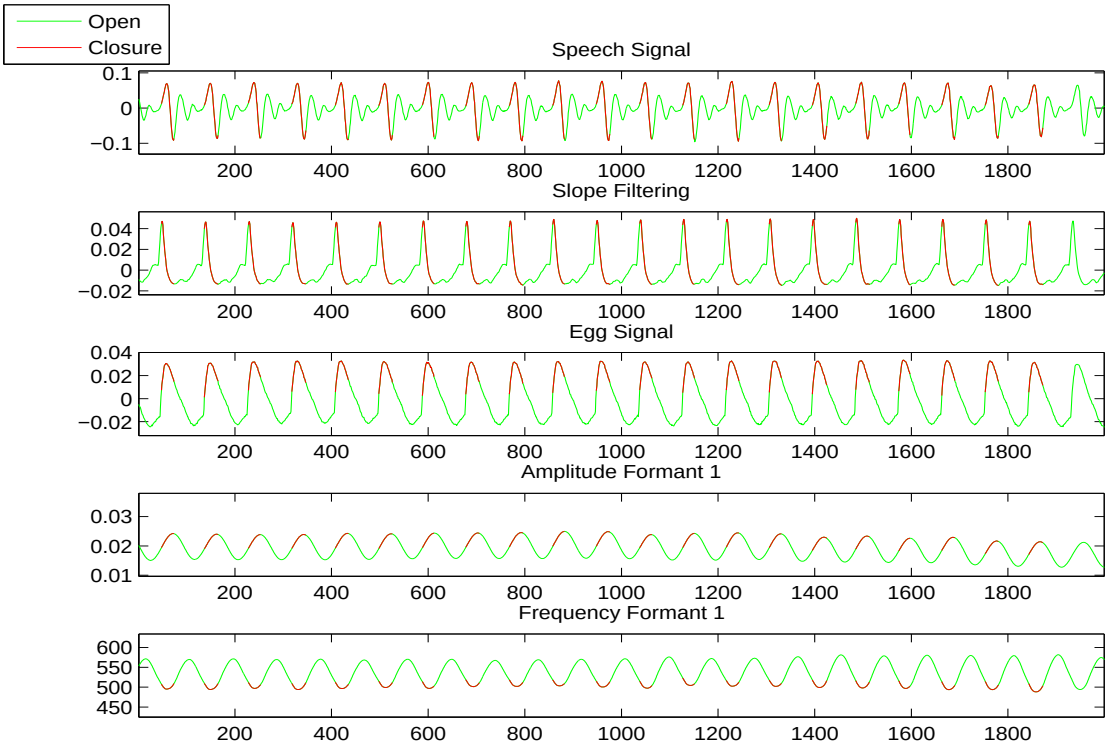



Figure 5.2: Recording of sustained vowel /o/ from Female Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 1st Formant, (e) Frequency Component of 1st Formant

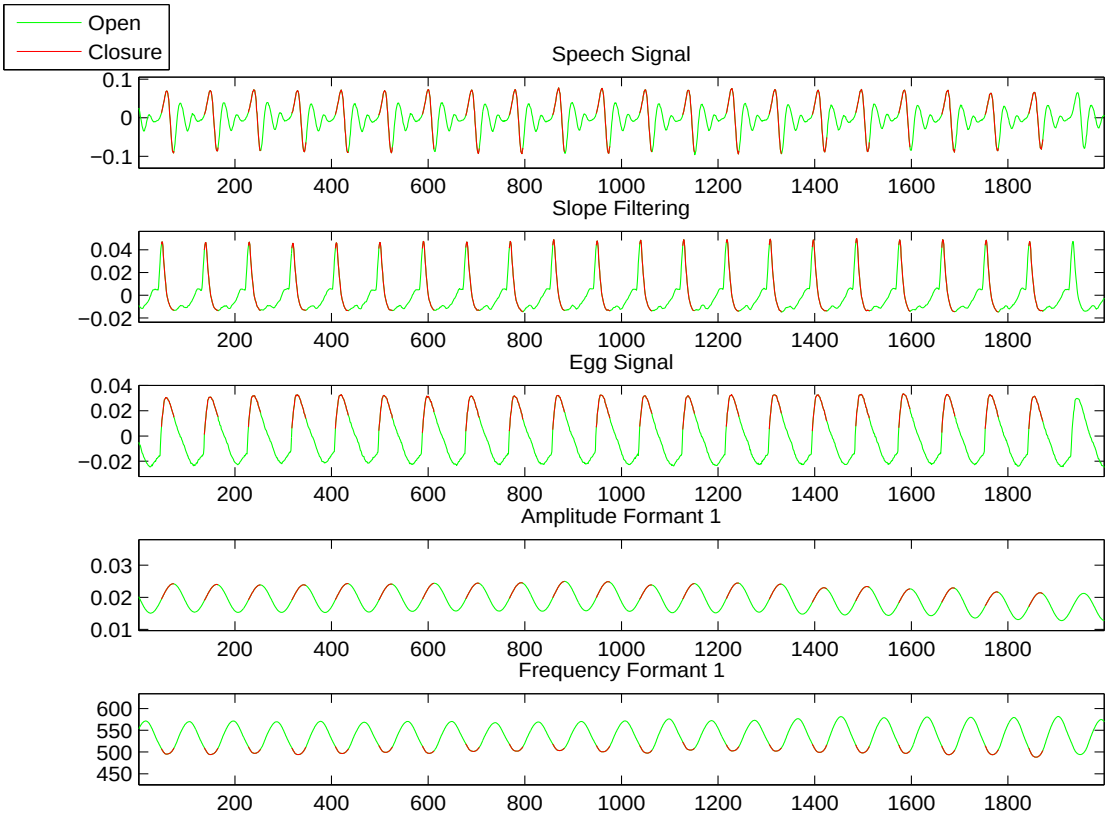



Figure 5.3: Recording of sustained vowel /o/ from Female Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 2nd Formant, (e) Frequency Component of 2nd Formant

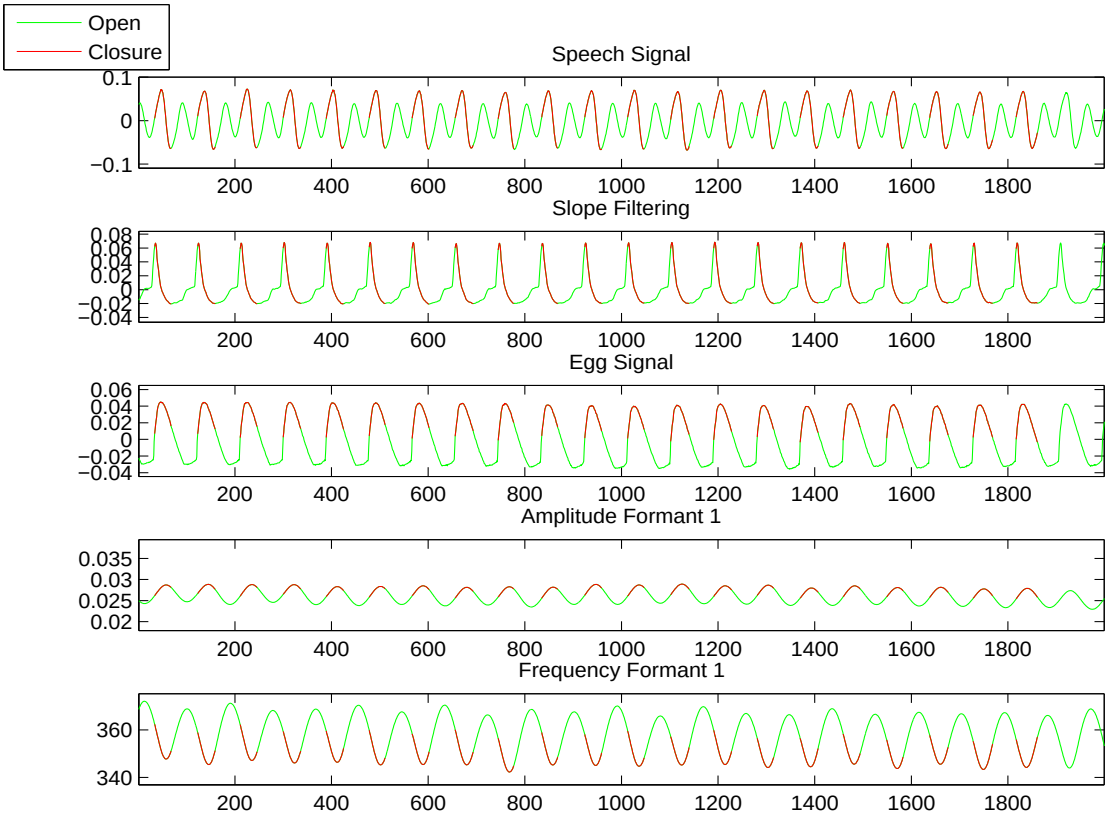



Figure 5.4: Recording of sustained vowel /u/ from Female Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 1st Formant, (e) Frequency Component of 1st Formant

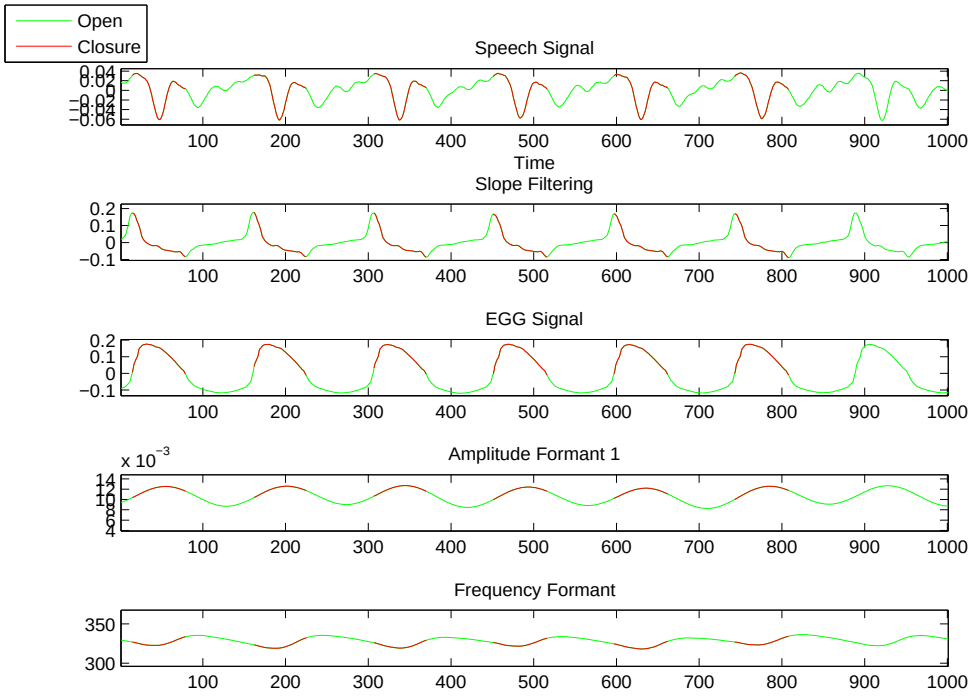



Figure 5.5: Recording of sustained vowel /u/ from Male Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 1st Formant, (e) Frequency Component of 1st Formant

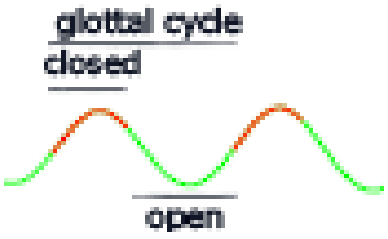



Figure 5.6: Closed and Open phase in a glottal cycle on the Amplitude Component of Formant

## 5.4 Discussion

This chapter has concentrated on the high resolution speech analysis. The glottal area function with the bandwidth and formant frequency of first $F1$ formant was discussed in 2.2 and illustrated in 2.3. The developed model verified that the increase in bandwidth is caused when the glottis opens and when it closes, the bandwidth drops. During the closed phase of glottis the bandwidth remains invariable and close to zero value. The formant Frequency rises at the onset of the glottal open phase and falls near the termination of the phase. The observations derived from the model are not expected to be completely represented in real speech signals. However, correlations between the components of speech signal and those of the model can be proposed.

The AM and FM components derived from the AM–FM Decomposition algorithm, which was described in this chapter, are observed during open and closed glottal phases. The glottal phases are generated from the EGG signal and are time–aligned with the speech signal. The results of the applied aQHM algorithm that presented in this chapter demonstrate that the frequency formant follows the descripted model in general. The amplitude of the formant increases at the open phase and decreases at the glottal closure instant. During the closed phase it is not obtained a steady value as in the model. The fact that the signal is real justifies this effect. On the other hand, the amplitude modulation of each formant is increased during glottal closure and falls in glottal opening phase.

Furthermore, based on the current results and observations a long Hamming window ($\approx$ 30ms) is proposed to be applied in aQHM algorithm to reveal the sinusoidal frequencies that are close to the First Formant. A quadratic interpolation will show the frequency and amplitude modulations of the formant. The purpose is the disclosure of any existent correlation of the Formant's modulation with the glottal phases. Moreover, it would be of great significance to observe the correlation among the amplitude and frequency activity of formant for different vowels.

# Appendix A

# Abbreviations

| | |
|---|---|
| AM | Amplitude Modulation |
| aQHM | adaptive Quasi–Harmonic Model |
| CL | Criterion Level |
| Cq | Closed quotient |
| DECOM | DEGG–Correlation based method for Qpen quotient Measurement |
| DECPA | Derivative EGG Closure Peak Amplitude |
| DFT | Discrete Fourier Transform |
| DEGG | Derivative of EGG |
| DYPSA | Dynamic Programming Projected Phase–Slope Algorithm |
| EGG | Electroglottograph |
| FAR | False Alarm Rate |
| FIR | Finite Impulse Response |
| FM | Frequency Modulation |
| GCI | Glottal Closure Instant |
| GOI | Glottal Opening Instant |
| IDA | Identification Accuracy |
| IDR | Identification Rate |
| LPC | Linear Prediction Coding |
| MR | Miss Rate |
| Oq | Open quotient |
| QHM | Quasi–Harmonic Model |
| SIGMA | Singularity detection In EGG with Multiscale Analysis |

| | |
|---|---|
| SNR | Signal–to–Noise Ratio |
| SPAR | Speech–Pattern Algorithms and Representations |
| SFS | Speech Filing System |
| SVD | Singular Value Decomposition |
| VFCA | Vocal Fold Contact Area |
| HQTx | High Quality Tx |

# Appendix B

# Amplitude-Frequency Modulations during Glottal Phases

The figures below depict the behavior of Amplitude and Frequency Components during the closed and open phase of vocal folds. Each Figure shows the recorded speech signal, the slope filtered EGG signal, the EGG signal and the AM, FM components derived from aQHM method (as described in Section 5.2). The GCIs, GOIs are extracted with the slope filtering process of time–aligned EGG signal. The closed and open phases are illustrated in all signals with red and green color respectively. The applied method is thoroughly explained in Section 5.3.

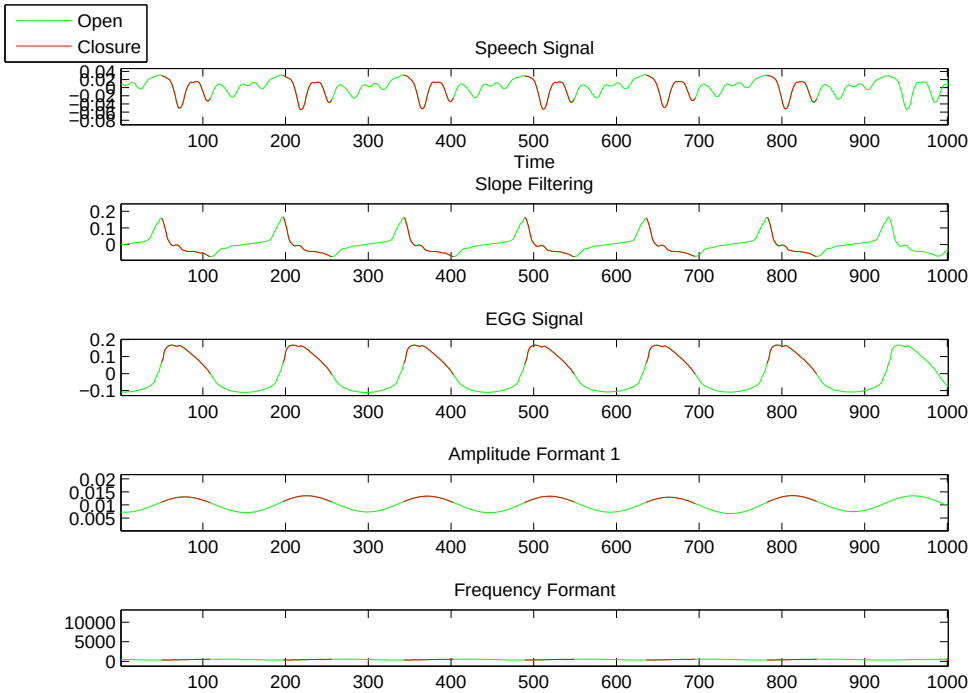



Figure B.1: Recording of sustained vowel /o/ from Male Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 1st Formant, (e) Frequency Component of 1st Formant

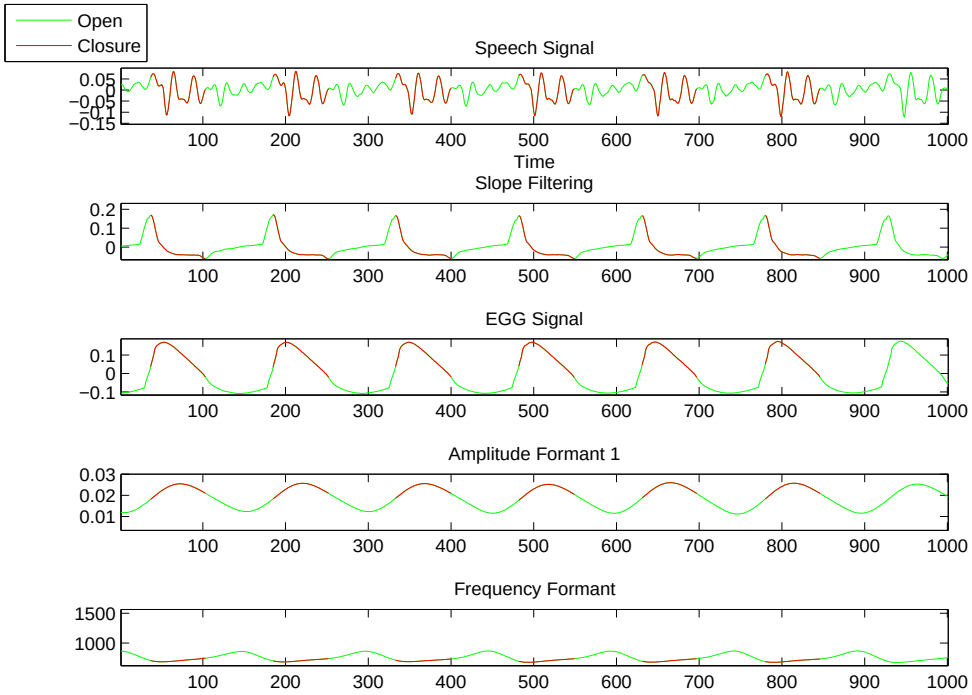



Figure B.2: Recording of sustained vowel /a/ from Male Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 1st Formant, (e) Frequency Component of 1st Formant

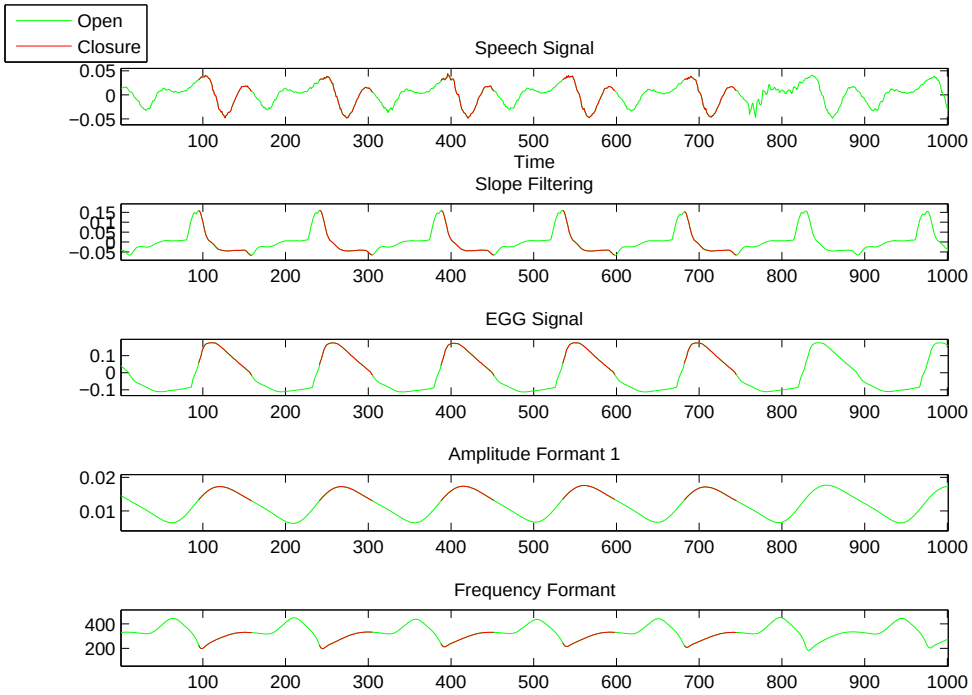



Figure B.3: Recording of sustained vowel /i/ from Male Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 1st Formant, (e) Frequency Component of 1st Formant

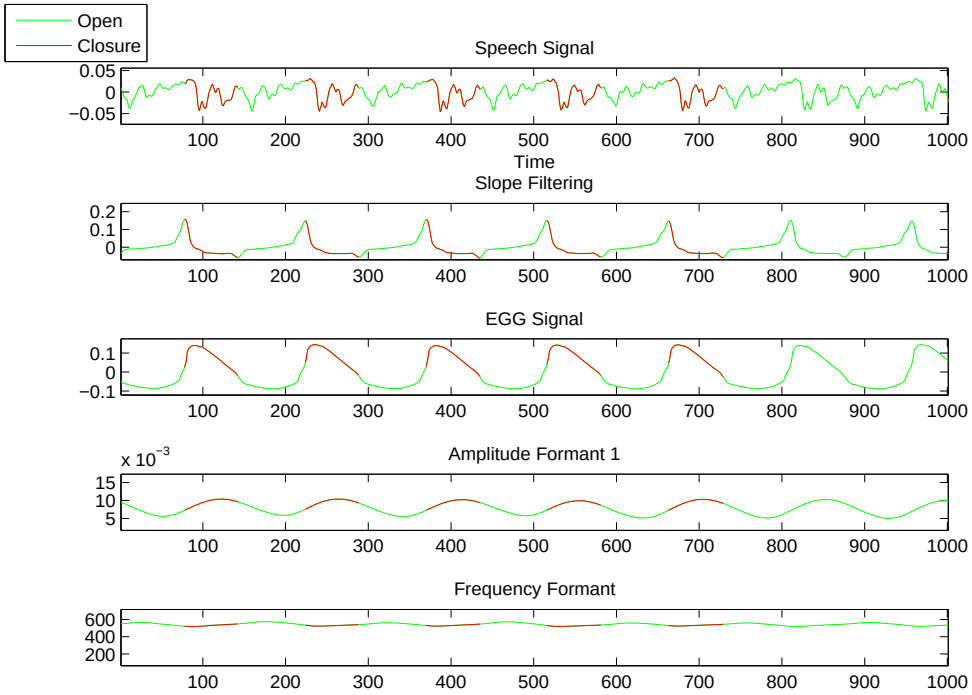



Figure B.4: Recording of sustained vowel /e/ from Male Speaker:(a) Speech Signal, (b) Slope Filtered EGG, (c) EGG Signal, (d) Amplitude Component of 1st Formant, (e) Frequency Component of 1st Formant

# References


[1] Abberton, E., R., M., Howard, D., M., and Fourcin, A. Laryngographic Assessment of Normal Voice: A Tutorial. *Clinical Linguistics and Phonetics*, 3:281–296, 1989.

[2] Ananthapadmanabha, T., V., and Yegnanarayana, B. Epoch Exctraction from Linear Prediction Residual for Identification of Closed Glottis Interval. *IEEE Transactions on Acoustic, Speech, Signal Processing*, ASSP–27(4):309–319, August 1979.

[3] Austin, S., F. and Titze, I., R. The Effect of Subglottal Resonance Upon Vocal Fold Vibration. *Journal of Voice*, 11(4):391–402, 1997. Lippincott–Raven Publishers.

[4] Bouzid, A. and Ellouze, N. Glottal Opening Instant Detection from Speech Signal. In *Proc. European Signal Processing Conference (EUSIPCO)*, pages 729–732, Vienna, September 2004.

[5] Bouzid, A., and Ellouze, N. Multiscale product of Electroglottogram Signal for Glottal Closure and Opening Instant Detection. *Proc. IMACS MultiConference on Computational Engineering in Systems Applications*, 1:106–109, 2006.

[6] Bouzid, A. and Ellouze, N. Open Quotient Measurements Based on Multiscale Product of Speech Signal Wavelet Transform. *Research Letters in Signal Processing*, 62521, 2007.

[7] Brookes, D., M., and Loke, H., P. Modeling Energy Flow in the Vocal Tract with Applications to Glottal Closure and Opening Detection. In *Proc. ICASSP*, pages 213–216, March 1999.

[8] Brookes, M. Glottal Closure Identification in Voiced Speech, Imperial College London. [Online PDF], February 2003. `http://mi.eng.cam.ac.uk/seminars/speech/brookes_seminar1.pdf`.

[9] Brookes, M., Naylor, P., A., and Gudnason, J. A Quantitative Assessment of Group Delay Methods for Identifying Glottal Closures in Voiced Speech. *IEEE Transactions on Audio, Speech, and Language Processing, IEEE Signal Processing Society*, 14(2):456–466, 2006.

[10] Carlson E. Accent method plus direct visual feedback of electroglottographic signals. *Voice Therapy, Clinical Studies*, 30(2):57–71, 1993. Mosby Year Book.

[11] Carlson E. Electrolaryngography in the assessment and treatment of incomplete mutation (puberphonia) in adults. *International Journal of Language and Communication Disorders*, 30(2):140–148, 1995.

[12] Carlson, E., and Miller, D. Aspects of Voice Quality: Display, Measurement and Therapy. *International Journal of Language and Communication Disorders*, 33:304–309, 1998.

[13] Childers, D., G., Hooks, D., M., Moore, G., P., Eskenazi, L., and Lalwani, A. I. Electroglottography and Vocal Fold Physiology. *Journal of Speech, Language and Hearing Research*, 33(2):245–254, 1990.

[14] Degottex, G., Roebel, A., and Rodet, X. Glottal Closure Detection from a Glottal Shape Estimate. In *SPECOM*, pages 345–349, St. Petersburg, June 2009.

[15] Drugman, T., and Dutoit, T. Glottal Closure and Opening Instant Detection from Speech Signals. *Proc. INTERSPEECH*, September 6–10 2009. Brighton UK.

[16] Fourcin, A. Laryngographic assessment of phonatory function. In C.L. Ludlow and M.O. Hart, editor, *Proc. of the Conference of the Assessment of Vocal Pathology*, volume 11, pages 116–127, Maryland, 1981.

[17] Fourcin, A. Electrolaryngographic assessment of vocal fold function. *Journal of Phonetics*, 14:435–442, 1982.

[18] Fourcin, A. *Voice Quality Measurement:Precision Stroboscopy, Voice Quality and Electrolaryngography*, chapter 13. Singular Publishing Group, San Diego, 2000.

[19] Fourcin, A., J., and Abberton, E. Laryngograph Studies of Vocal Fold Vibration. *Phonetica*, 34:313–315, 1977.

[20] Fourcin, A., McGlashan, J., and Blowes R. Measuring Voice in the Clinic–Laryngograph Speech Studio Analyses. In *6th Voice Symposium of Australia*, Adelaide, October 2002.

[21] Greater Baltimore Medical Center. `http://www.gbmc.org/`.

[22] Henrich, N., d' Alessandro, C., Doval, B., and Castellengo, M. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Acoustical Society of America*, 115(3):1321–1332, 2004.

[23] Henrich, N., Roubeau, B., and Castellengo, M. On the Use of Electroglottography for Characterisation of the Laryngeal Mechanisms. In *Proc. of the Stockholm Music Acoustics Conference (SMAC)*, Stockholm, Sweden, August 6–9 2003.

[24] Höge, H., Kotnik, B., Kacic, Z., and Pfitzinger, H., R. Evaluation of Pitch Marking Algorithms. In *Proc. ITG Fachtagung Sprach Kommunikation*, Kiel, Germany, 2006.

[25] Howard, D., M. Variation of Electrolaryngographically derived Closed Quotient for trained and untrained Adult Female Singers. *Journal of Voice*, 9(2):163–172, June 1995.

[26] Howard, D., M., Lindsey, G., A., and Allen, B. Toward the Quantification of Vocal Efficiency. *Journal of Voice*, 4(3):205–212, 1990. Raven Press, Ltd.

[27] Huckvale, M. Speech Filing System: Tools for speech (SFS). Technical report, University College London, 2008. `http:www.phon.ucl.ac.uk/resource/sfs/`.

[28] Huckvale, M., A., Brookes, D., M., Dworking, L., T., Johnson, M., E., Pearce, D., J., Whitaker, L. The SPAR Speech Filing System. In *European Conference on Speech Technology*, pages 305–308, Edinburgh, 1987.

[29] Klatt, D., H. and Klatt, L. C. Analysis, Synthesis and Perception of Voice Quality variations among female and male talkers. *Acoustical Society of America*, 87(2):820–857, February 1990.

[30] Kounoudes, A., Naylor, P., A., and Brookes, M. Automatic Epoch Extraction for Closed–Phase Analysis of Speech. In *14th International Conference on Digital Signal Processing Proceedings (DSP)*, pages 979–983, 2002.

[31] Kounoudes, A., Naylor, P., and Brookes, M. The DYPSA Algorithm for Estimation of Glottal Closure Instants in Voiced Speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 15, pages 349–352, 2002.

[32] Laryngograph, Ltd. No1. Foundry Mews, London NW1 2PR United Kingdom. `http://www.laryngograph.com/`.

[33] Ma, C., Kamp, Y., and Willems, L., F. A Frobenius Norm approach to Glottal Closure Detection from the Speech Signal. *IEEE Transcations on Speech Audio Processing*, 2(2):258–265, April 1994.

[34] Maqsood, H., Gudnason, J., and Naylor, P., A. Enhanced Robustness to Unvoiced Speech and Noise in the DYPSA Algorithm for Identification of Glottal Closure Instants. In *Proc. European Signal Processing Conference(EUSIPCO)*, pages 2310–2315, 2007.

[35] Maqsood, H., Patrick, D., and Naylor, A. Improved DYPSA Algorithm for Noise and Unvoiced Speech. In *International Conference on Emerging Technologies (ICET)*, pages 243–248, Islamabad, 2007.

[36] Marasek, K. Egg and Voice Quality. Technical report, Institute for Natural Language Processing, Universitat Stuttgart, 1997. `http://www.ims.uni-stuttgart.de/phonetik/`.

[37] McKenna, J., G. Automatic glottal closed–phase location and analysis by Kalman filtering. In *4th ISCA Tutorial and Research Workshop on Speech Synthesis, Pittlochrie*, Scotland, August 2001.

[38] Michaud, A. A Measurement from Electroglottography: DECPA, and tis Application in Prosody. In *International Conference Speech Prosody*, Nara, Japan, March 23–26 2004.

[39] Naylor, P., A., Kounoudes, A., Gudnason, J., and Brookes, M. Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm. *IEEE Transactions on Audio, Speech, and Language Processing, IEEE Signal Processing Society*, 15(1):34–43, 2007. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/dypsa.html`.

[40] Pantazis, Y., Rosec, O., and Stylianou, Y. On the Properties of a Time–Varying Quasi–Harmonic Model of Speech. *INTERSPEECH*, pages 1044–1047, September 2008.

[41] Quatieri, T., F. *Discrete–Time Speech Signal Processing: Principles and Practice*. Alan V. Oppenheim, Series Editor, 2002. Prentice Hall.

[42] Rothenberg, M., and Mashie, J., J. Monitoring Vocal Fold Abduction Through Vocal Fold Contact Area. *Journal of Speech and Hearing Research*, 31:338–351, September 1988.

[43] Saidi, W., Bouzid, A., and Ellouze, N. Evaluation of Multi–scale Product Method and DYPSA Algorithm for Glottal Closure Instant Detection. In *Proc. 3rd International Conference on Information and Communication Technologies: From Theory to Applications (ICTTA)*, pages 1–5, April 2008.

[44] Schnell, K. *Advances in Nonlinear Speech Processing:Estimation of Glottal Closure Instances from Speech Signals by Weighted Nonlinear Prediction*, volume 4885, pages 221–229. Springer Berlin/Heidelberg, 2007.

[45] Smits, R., and Yegnanarayana, B. Determination of Instants of Significant Excitation in Speech Using Group Delay Function. *IEEE Transactions on Speech Audio Processing*, 3:325–333, 1995.

[46] Steiglitz, K., and Dickinson, B. The Use of time–domain selection for Improved Linear Prediction. *IEEE Transcations on Acoustic, Speech, Signal Processing*, ASSP–25(1):34–39, February 1977.

[47] Strube, H., W. Determination of the Instant of Glottal Closures from the Speech Wave. *Acoustical Society of America*, 56(5):1625–1629, November 1974.

[48] Stylianou, Y. *Harmonic plus Noise Models for Speech, combines with Statistical Methods*. PhD thesis, Ecole Nationale Supérrieure des Télécommunications, 1996.

[49] Stylianou, Y. Synchronization of Speech Frames based on Phase Data with Application to Concatenative Speech Synthesis. In *Proceedings 6th European Conference Speech Communication and Technology, EUROSPEECH*, volume 5, pages 2343–2346, Budapest, Hungary, September 5–9 1999.

[50] Thomas, M., R., P., and Naylor, P., A. The SIGMA Algorithm: A Glottal Activity Detector for Electroglottographic Signals. *IEEE Transactions on Audio, Speech, and Language Processing, IEEE Signal Processing Society*, 6(1), January 2007.

[51] Thomas, M., R., P., and Naylor, P. A. The SIGMA Algorithm for Estimation of Reference–Quality Glottal Closure Instants from Electroglottograph Signals. In *Proc. European Signal Processing Conference (EUSIPCO))*, Lausanne, Switzerland, August 2008.

[52] Thomas, M., R., P., Gudnason J. and Naylor, P. A. Application of the DYPSA algorithm to segmented time-scale modification of speech. In *Proc. European Signal Processing Conference (EUSIPCO))*, Lausanne, Switzerland, August 2008.

[53] Trefethen, L., N. *Spectral Methods in MATLAB*. SIAM, Philadelphia, 2000.

[54] Tuan, V., N., and d'Alessandro, C. Robust Glottal Closure Detection using Wavelet transform. In *Proc. EUROSPEECH*, pages 805–808, Budapest, September 1999.

[55] Turner, C., S. Slope Filtering: An FIR Approach to Linear Regression. *IEEE Signal Processing Magazine, DSP Tips and Tricks*, pages 159–163, November 2008.

[56] Wong, D., Y., Markel, J., D., and Gray, A., H. Least Squares Inverse Filtering from the Acoustic Speech Waveform. *IEEE Transcations on Acoustic, Speech, Signal Processing*, ASSP–27(4):350–355, 1979.