Speech Production Modelling and Analysis

Andreas I. Koutrouvelis

Master of Science Thesis

Committee: Dr.ir. R. Heusdens (Supervisor)
Prof.dr.ir. A.J. van der Veen
Dr.ir. J. Weber
Dr. N.D. Gaubitch

Delft
University of
Technology

Speech Production Modelling and Analysis

Master of Science Thesis

For the degree of Master of Science in Electrical Engineering at Delft University of Technology

Andreas I. Koutrouvelis

July 16, 2014

Faculty of Electrical Engineering, Computer Science and Mathematics (EEMCS) · Delft University of Technology

Abstract

The first part of the present thesis reviews the speech production mechanism and several models of the glottal flow derivative waveform and of the vocal tract filter. The source filter model is investigated in depth, since it is the most important "ingredient" of linear prediction analysis. We also review seven linear prediction (LP) methods based on the same general LP optimization framework. Moreover, we examine the importance of pre-emphasis and glottal-cancellation prior to LP.

The second part of the thesis, provides an experimental evaluation of the LP methods combined with several pre-emphasis and glottal-cancellation techniques in the context of two general application areas. The first area consists of applications which aim to estimate the true glottal flow or glottal flow derivative signal. The second area consists of applications which aim to find a sparse residual. In particular, five factors are investigated: the sparsity of the residual using the Gini index, the estimation accuracy of the glottal flow derivative using the signal to noise ratio (SNR), the estimation accuracy of the vocal tract spectral magnitude using the log spectral distortion distance (LSD) metric, and the probability of obtaining a stable linear prediction filter. All these factors are evaluated for clean and reverberated speech signals. The sparse linear prediction methods and the iteratively reweighted least squares method combined with a second order pre-emphasis filter give the most accurate glottal flow derivative estimates, the most accurate vocal tract estimates and the sparsest residuals in most cases. Finally, we compare several linear prediction methods in the context of the speech dereverberation method proposed in [1, 2]. This method enhances the reverberated residual obtained via the autocorrelation method. In the context of this application, we show that the sparse linear prediction method and the weighted linear prediction method combined with a second-order pre-emphasis filter perform better than the autocorrelation method.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I wish to express my heartfelt appreciation and gratitude to my supervisor, Associate Professor Richard Heusdens, who assigned to me the present task. His suggestions, criticism, personal guidance and his course Audio and Speech Processing were invaluable in understanding the important issues related to the topic of the thesis and the culmination of this work. I am also grateful to Dr. Nikolay Gaubitch who helped me to understand some parts of the theory and provided me with useful information and suggestions. Many thanks also to Dr. George Kafentzis, from the University of Crete, with whom I discussed many aspects of speech signal processing. His suggestions and help were very important for accomplishing my thesis. Moreover, many thanks to Professor Brad Story from the University of Arizona, who has helped me to understand the software LeTalker and has provided useful code to me.

Finally, I wish to thank my parents Ioannis and Eleni, my sister Fani and my good friends for their support and encouragement during all my student life at Delft University of Technology.

Delft, University of Technology   Andreas I. Koutrouvelis
July 16, 2014

To the memory of my uncle Christos.

Chapter 1

Introduction

According to the source filter model (SFM), a short-time segment of a speech signal, $x[n]$; $n = 0, ..., N-1$, can be generated as the convolution of a source signal, $e[n]$, and a filter, $h[n]$ [3, 4]. A speech segment can be classified as voiced or unvoiced. In the voiced case, the source signal can be either the glottal flow derivative waveform, $\dot{u}_g[n]$, or a quasi-periodic impulse train signal, $p[n]$. In unvoiced case the source is a random noise signal which can be approximated by a white Gaussian noise (WGN) process. The glottal flow derivative signal can be generated as the convolution of the periodic impulse train signal, $p[n]$, the glottal pulse, $g[n]$, and the filter of the lips, $r[n]$. The source is considered as the impulse train, $p[n]$, only if the glottal pulse and the filter of the lips are considered parts of $h[n]$. The filter $h[n]$, without including $r[n]$ and $g[n]$, is the vocal tract filter which is usually modeled as an all-pole filter [4]. Therefore, based on the assumptions of SFM and of the all-pole structure of $h[n]$, a speech segment can be written using an auto-regressive model [4]

$$x[n] = \sum_{k=1}^{q} a_k x[n-k] + e[n], \tag{1-1}$$

where a_k, $k = 1, ..., q$, are the filter coefficients, called linear prediction coefficients (LPCs). Note that, when the source is assumed to be the glottal flow derivative signal, q is set approximately to the true order of the vocal tract filter. On the other hand, when the source is assumed to be the periodic impulse train, q is set to a much higher value. This is because, in this case, the filter, $h[n]$, consists of the vocal tract filter the lips filter and the glottal pulse.

The main purpose of linear prediction (LP) analysis is two estimate the source and/or the filter. In particular the LP analysis methods estimate the LPCs such that $e[n]$ is minimized. The resulting minimum $e[n]$, say $\hat{e}[n]$, is called the residual. The classical LP analysis method minimizes the variance (i.e., the squared L_2 norm) of $e[n]$ and, therefore, results in a linear least squares estimator (LLSE) [5]. The LLSE estimator works well for unvoiced speech, and it is equivalent to the maximum likelihood estimator MLE if $e[n]$ is indeed WGN [6]. On the other hand, in voiced speech the estimation of the source is more challenging since it is not WGN and the LLSE estimator does not perform very well.

There are *two general problems* in speech analysis which concern us in the present thesis. The first problem aims to estimate the true glottal flow derivative signal. Therefore, in this

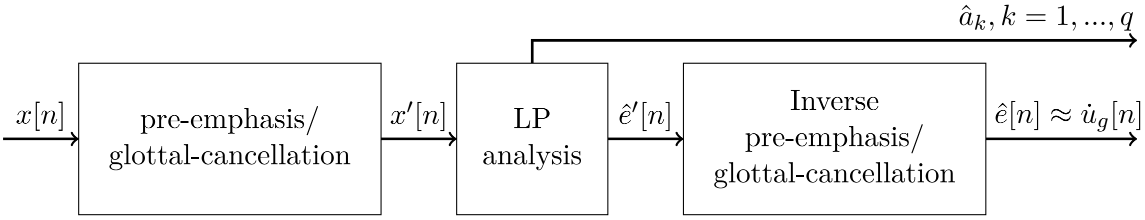

$x[n]$
pre-emphasis/ glottal-cancellation
$x'[n]$
LP analysis
$\hat{e}'[n]$
Inverse pre-emphasis/ glottal-cancellation
$\hat{e}[n] \approx \dot{u}_g[n]$
$\hat{a}_k, k = 1, ..., q$

Figure 1-1: Estimation of the glottal flow derivative $\dot{u}_g[n]$.

case, the source is assumed to be the glottal flow derivative signal. In this problem, the general speech analysis framework consists of three parts (see Figure 1-1). The first part is a pre-emphasis or a glottal-cancellation filter applied to the speech signal prior to LP analysis. The second part is LP analysis which estimates the filter coefficients and the pre-emphasized or glottal-canceled version of the glottal flow derivative, $e'[n]$, which is the convolution of the glottal flow derivative with the pre-emphasis or glottal-cancellation filter. The third part consists of the inverse pre-emphasis or glottal-cancellation filter, which estimates the glottal flow derivative signal. The estimation and modeling of the glottal flow derivative signal is important in many applications such as speech synthesis [7,8], LP speech coding [9], analysis of vocal emotions [10,11], analysis of pathological voices [12] or speaker identification [13]. For instance, the speaker identification application is based on the fact that the structure of the glottal flow derivative signal may contain several characteristics which are "unique" for each speaker. Moreover, the identification of several pathological voices, such as diplophonia and vocal fry is based on the structure of the glottal flow derivative signal (see more in Chapter 2).

The second general speech analysis problem aims to find a sparse residual. Therefore, the source is assumed to be the quasi-periodic impulse train signal. In this case, the general speech analysis framework typically consists of only the LP analysis part (see Figure 1-2). Nevertheless, in the present thesis we will show that pre-emphasis and glottal-cancellation increases sparsity and, therefore, we are going to test both schemes for this problem. A sparse residual is important in LP speech coding [4], speech enhancement / dereverberation [1, 2], epoch extraction [14] or speaker localization [15]. For instance, the performance of the speaker localization application is dependent on how strong are the main epochs of the residual compared to all the other values. These main epochs are the impulses of the impulse train signal.

The source signal in the first application area and the source signal in the second application area combined with a pre-emphasis filter or a glottal-cancellation filter, consist of quasi-periodic strong peaks and the LLSE suffers from outliers, i.e., it overemphasizes the large errors and puts less emphasis on smaller errors [5], producing a non-spiky residual. The

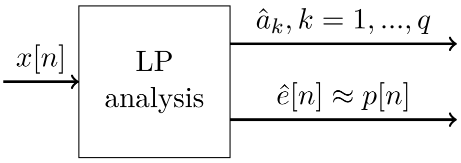


Figure 1-2: Estimation of the quasi-periodic impulse train $p[n]$.

property of spikiness on just a few samples is also called sparsity. Thus, in case of voiced speech, a desired property for an LP estimator is to estimate the LPCs such that the residual is sparse. The methods that follow this philosophy are called sparse LP methods.

1-1 Goals of the Thesis

As was explained previously, speech analysis is an easy task for the unvoiced speech. On the other hand, it is much more difficult for voiced speech. Therefore, the present thesis examines only the latter case. There are several LP methods in the literature trying to solve the two aforementioned general speech analysis problems. Some of those LP methods have not been tested in both problems. Moreover, to the author's knowledge, the second-order pre-emphasis and the glottal-cancellation techniques have not been used in the context of finding a sparse residual. Moreover, in LP speech coding applications it is important for a LP method to produce as less as possible unstable filters, while in glottal flow derivative estimation problems the contrary may be convenient. Therefore, the main objective of the present thesis is to answer the following three general questions.

1. Which LP method gives the sparsest residual, and how robust it is in reverberation phenomena? Do pre-emphasis and glottal-cancellation increase the sparsity of the residual?
2. Which LP method gives the most accurate glottal flow derivative and vocal tract estimates, and how robust it is when the speech signal is subject to reverberation? Which is the best pre-emphasis or glottal-cancellation technique in the context of this problem?
3. What is the probability of estimating an unstable filter in both general analysis problems?

1-2 Contribution and Previous Work

The main contribution of the present thesis is an experimental evaluation of seven LP methods, combined with several pre-emphasis and glottal-cancellation filters, in the context of the two aforementioned application areas. In particular, five factors are evaluated: the sparsity of the residual using the Gini index [16], the estimation accuracy of the glottal flow derivative using the signal to noise ratio (SNR), the estimation accuracy of the vocal tract spectral magnitude using the log spectral distortion distance (LSD) metric [17], and the probability of obtaining a stable LP filter. All these factors are evaluated for clean and reverberated speech signals. Through this experimental evaluation, we empirically show the following.

1. The glottal-cancellation and the pre-emphasis filters increase the sparsity of the residual, when the LP order is approximately set to the true order of the vocal tract filter.
2. The probability of obtaining a stable filter is not decreased due to reverberation phenomena. On the contrary, in some cases it increases slightly. Moreover, when pre-emphasis and glottal-cancellation filters are used, the stability increases.
3. The sparse linear prediction methods [18,19] and the iteratively reweighted least squares method [20] combined with a second-order pre-emphasis filter are the most accurate estimation methods of the glottal flow derivative and the vocal tract filter. Their performance is also high when they are applied on reverberated speech, but in this case the weighted linear prediction method [21] combined with a second-order pre-emphasis

filter performs slightly better. A popular glottal flow estimation method, called iterative adaptive inverse filtering (IAIF) [22,23], is compared with the three aforementioned combinations and it is found less accurate, especially when the glottal formant is close to the first formant of the vocal tract.

4. The sparse linear prediction methods and the iteratively reweighted least squares method combined with a second-order pre-emphasis filter, give the sparsest residuals.

Moreover, we highlighted several possible problems of LP analysis that may be solved by using a pre-emphasis or a glottal-cancellation filter prior to LP. To the author's knowledge, there has not been any previous document which gathers and lists all these problems. During our effort to list all these problems, we found that pre-emphasis and glottal-cancellation techniques increase the region where the glottal flow derivative is zero and, therefore, a longer analysis interval can be taken in the closed phase analysis method [24]. Our experimental evaluation empirically shows that this is true.

We should not forget to mention that the used weight function, in the iteratively reweighted least squares method, is the Andrew's function [25]. To the author's knowledge, this weight function has not been used for speech analysis before. We selected this function after testing several other weight functions presented in [20] which give less sparse residuals. This comparison is not included in the present thesis.

Finally, we compare several best performing LP methods in the context of the speech dereverberation method proposed in [1, 2]. This speech dereverberation method consists of two steps: the enhancement of the reverberated residual and the enhancement of the reverberated LP coefficients. In the present thesis, we managed to improve the enhancement of the reverberated residual by using the sparse prediction methods and the weighted linear prediction method instead of the autocorrelation method which was used in [1,2].

Based on part of the work done in this thesis, a paper co-authored by R. Heusdens and N.D. Gaubitch was presented in the 35th WIC Symposium on Information Theory in the Benelux, Eindhoven [26].

There are some previous experimental evaluations of several LP methods. In particular, in [27], the authors compared several LP methods, including three of the seven LP methods presented in the present thesis. These methods were compared in terms of the estimation accuracy of the speech spectral envelope. Note that the speech spectral envelope is not the same as the vocal tract spectral magnitude, as it is shown in Chapter 2. However, they did not use any pre-emphasis or glottal-cancellation in their paper. Moreover, in [28], several LP methods, including three of the seven methods used in the present thesis, were compared in terms of the estimation accuracy of the glottal flow signal. Unlike in [27], the author of [28] used a first-order pre-emphasis filter. However, we show in the present thesis that the second-order pre-emphasis filter and the glottal-cancellation filters perform better than the first-order pre-emphasis filter.

1-3 Outline

In Chapter 2, the fundamental mechanisms of speech production are presented. In particular, several models of the glottal flow derivative and their properties are presented. Two simple models are the Liljencrants-Fant (LF) model [29] and the Rosenberg model [30], and one more complicated mechanical model is the lumped-element model [31] which simulates the movements of the vocal folds. More emphasis is given to the LF model which is used in our experiments in Chapter 4. Furthermore, we present two different vocal tract models; the Kelly-Lochbaum model [32] and the Story model [33]. The first does not model the losses of the vocal tract while the second one does. The latter model is used in the experiments of Chapter 4, for the reconstruction of the vocal tract filter from a finite number of area functions acquired via MRI imaging [34]. Finally, we present the source filter model (SFM) [3] which is a simplification of the speech production mechanism. In Chapter 3, we see what is the relationship between SFM and LP and what are the consequences in the estimation accuracy of the LP methods, due to the simplifications that are introduced by SFM.

In Chapter 3, we review seven LP methods and their properties. There are two main categories of LP methods; those which are based on L_2 minimization and those which are based on L_1 minimization. Both categories are special cases of a general LP optimization problem [27]. Moreover, we investigate the importance of pre-emphasis and glottal-cancellation prior to LP. Specifically, a first-order and a second-order pre-emphasis filter are examined. In Chapter 4, we see that the latter outperforms to the first-order pre-emphasis filter in both applications areas. Furthermore, two glottal-cancellation methods are evaluated. The first glottal-cancellation method is the first part of the famous IAIF method [22, 23] and we call it IAIFGC. The second glottal-cancellation method is introduced in the present thesis, named SGPC. In Chapter 4, although we managed to improve the performance of IAIF by replacing its first part, IAIFGC, with SGPC, we found that there are other more accurate glottal flow derivative estimation methods with less complexity. Finally, in Chapter 4, the experimental evaluation explained in Section 1-2 is undertaken.

Chapter 2

Speech Production and Modeling

In this chapter, we review the anatomy of the human speech production system and its discrete-time/space realization, named speech production model or source filter model. Special emphasis is given to the modeling of the glottal flow, the glottal flow derivative and the vocal tract filter.

2-1 Anatomy of Speech Production

The main parts of the human body, responsible for the speech production (Figure 2-1) are: the *lungs*; the *trachea*, also known as the *subglottal area*; the *larynx*; and the *vocal tract*, also known as *supraglottal area* [3, 4, 35, 36]. The lungs behave like a power generator supplying the larynx with air. The major component of the larynx is a pair of *vocal folds/vocal chords*. The orifice of the vocal folds is called *glottis*. During speech, air is flowing from the lungs towards the vocal folds and the output of the vocal folds is a *time-varying velocity signal* called *glottal source* or simply *source*. During *unvoiced speech* the vocal folds remain open, while during *voiced speech* they oscillate with a certain frequency called *pitch* whose inverse is the pitch period. The source signal during voiced speech is called *glottal flow* and the glottal flow for one pitch period is called *glottal pulse*. A glottal pulse (Figure 2-2) is separated into three time regions: the *open phase*, the *return phase* and the *closed phase*, where the vocal folds are opening, closing, and stay closed, respectively. However, in reality, we do not have this ideal shape of the glottal flow as in Figure 2-2. The vocal folds of some people do not close completely during the closed phase, while others have vocal disorders such as *vocal fry* or *diplophonia* which add a secondary glottal pulse, in each pitch period, to the glottal flow signal (Figure 2-3) [4]. In case of *aspirated voicing* a part of the glottis remains slightly opened causing turbulence in the glottal flow [4, 13]. Moreover, the glottal flow structure is effected by its non-linear coupling with the subglottal and supraglottal areas [37]. Therefore, the modeling of the glottal flow is a difficult task because of the multi-varying nature.

The purpose of the vocal tract is to shape the source signal into perceptually speech sounds [4, 36]. It consists of five main parts: the *pharynx*, the *oral cavity*, the *nasal cavity*, the *velum* and the *lips*. The velum works as a three state switch selecting one of the two cavities or both

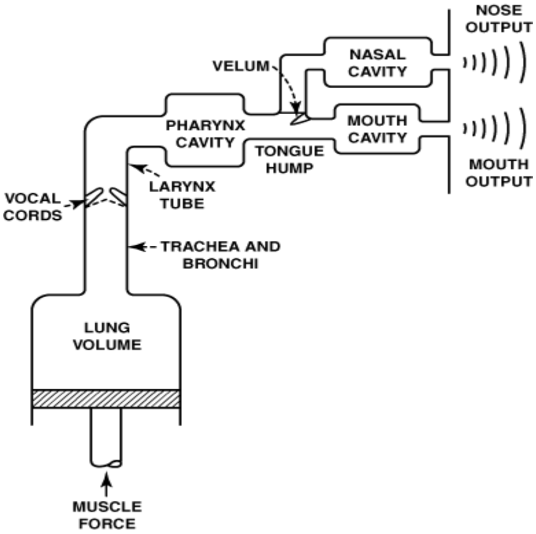


Figure 2-1: Speech Production System.

SOURCE: [36].

of them. When it is closed the oral cavity is used only contributing with *resonances/formants*. On the other hand, when the velum is open or being in the intermediate state, the nasal cavity contributes with *anti-resonances* and *nasalized speech* is produced. The pharynx combined with the oral cavity or/and the nasal cavity can be considered as an *acoustic filter* which takes as input the source signal and produces as output another velocity signal "colored" by those resonances and anti-resonances. When the oral cavity works alone, this acoustic filter can be discretized and approximated very well by the concatenation of a few *cylindrical tubes* (Figure 2-4) [4]. This approximation is very useful because it gives us an *all-pole minimum-phase linear filter*, where the filter coefficients are dependent on the current shape of the

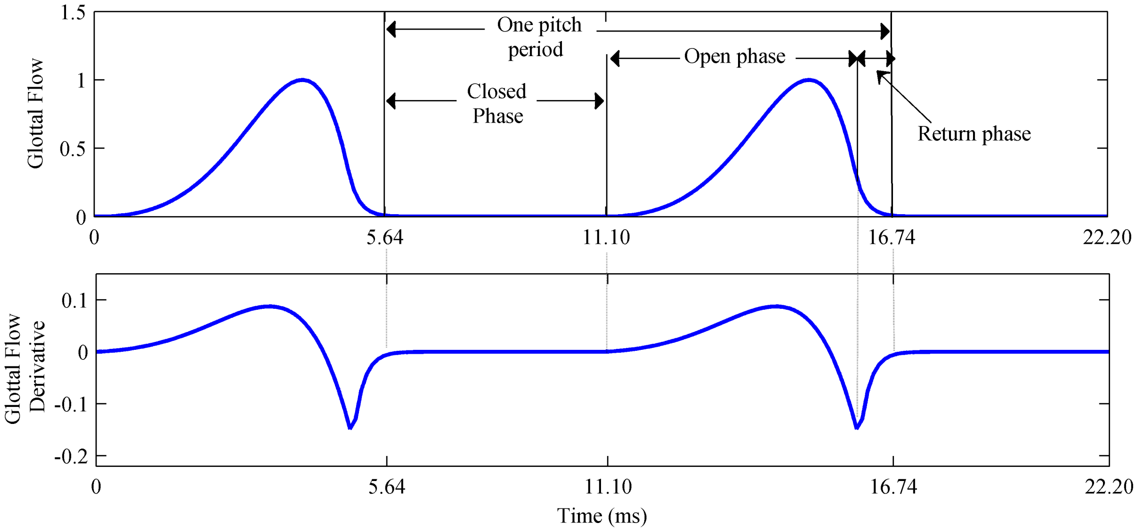


Figure 2-2: Two pitch periods of the glottal flow and the glottal flow derivative of a male speaker.

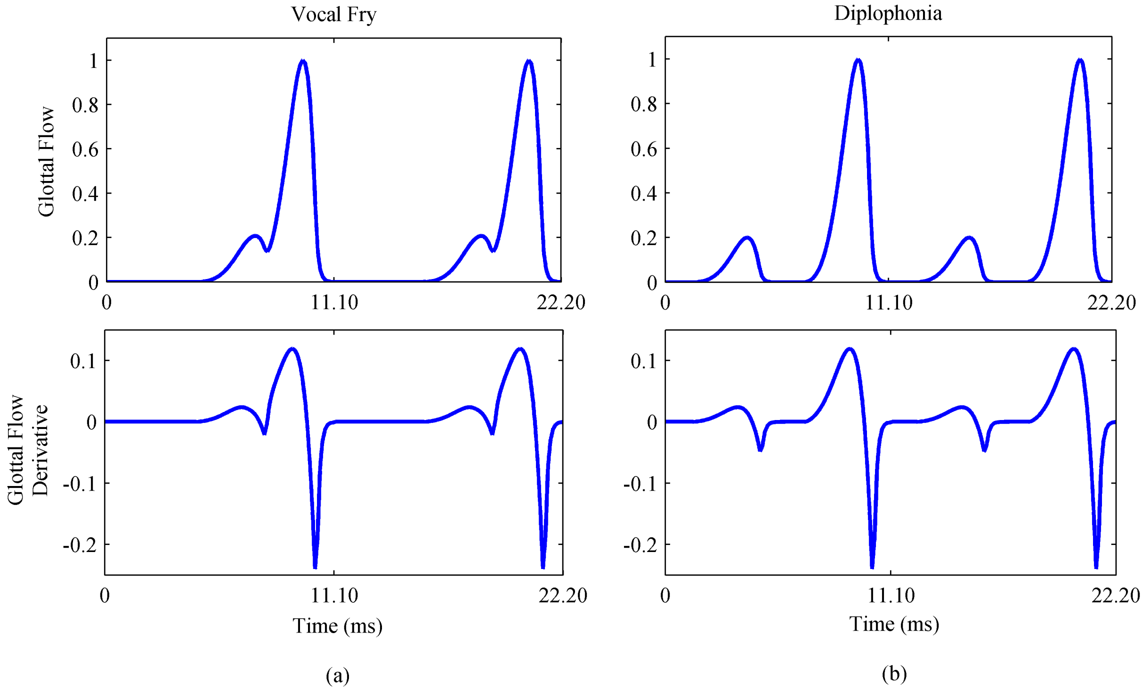


Figure 2-3: Two pitch periods of the glottal flow and the glottal flow derivative of a male speaker having vocal fry (a) and diplophonia (b).

acoustic filter. On the other hand, when the nasal cavity works alone or in conjunction with the oral cavity the acoustic filter has also anti-resonances and, therefore, it is better modeled with a *pole-zero linear filter* [4].

Finally, the lips work as a *differentiator* of the output velocity signal of the acoustic filter, converting it into a *pressure signal*, i.e., the speech we hear [4]. As we will see in Section 2-4, we can change the position of the transfer function of the lips and combine it with the source signal. This combination results in the *source derivative signal* and in the case of voiced speech, it is called *glottal flow derivative* (Figure 2-2). In the present thesis the combination of the lips with the glottal flow is used very often and, in order to avoid confusion in the sequel when we will refer to the vocal tract, we will mean the linear acoustic filter *without* the lips.

2-2 Modeling of the Glottal Flow and the Glottal Flow Derivative

The estimation and modeling of the glottal flow or the glottal flow derivative is very important in many applications such as speech synthesis [7, 8], speech coding [9], analysis of vocal emotions [10,11], analysis of pathological voices [12] or speaker identification [13]. A simplistic expression of the glottal flow signal $u_g[n]$ for a particular vowel with a constant pitch period is given by

$$u_g[n] = g[n] * p[n] = g[n] * \sum_{i=1}^{m} \delta[n - iT] = \sum_{i=1}^{m} g[n - iT], \tag{2-1}$$

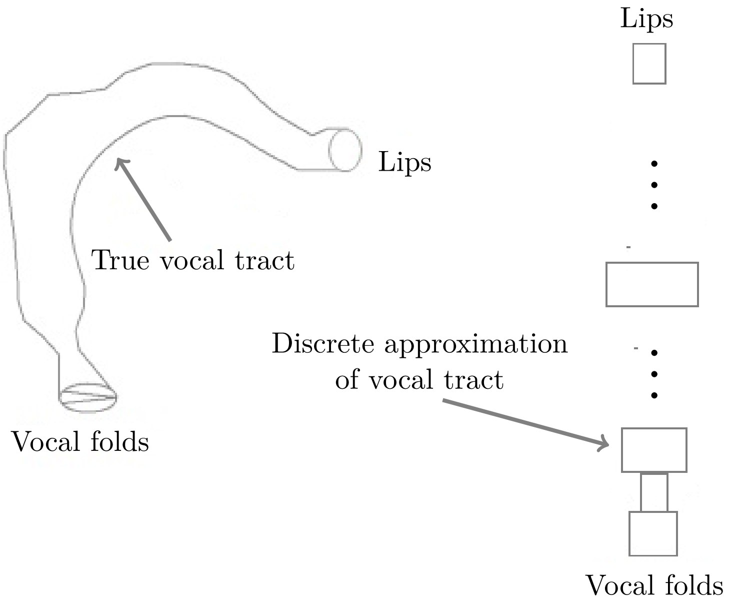


Figure 2-4: True vocal tract and its discrete approximation. The latter was made by hand.

SOURCE: The left part of the figure was taken from [38].

where $p[n]$ is a *periodic impulse train* signal, $g[n]$ is one glottal pulse, T is the pitch period and m is the number of pitch periods. The lips (i.e., the radiation impedance of the lips) are usually approximated by a high-pass finite impulse response (FIR) filter $r[n]$ with a minimum-phase transfer function having one zero inside and very close to the unit circle [4, 24, 39, 40], i.e.,

$$R(z) = 1 - \alpha z^{-1}, \tag{2-2}$$

where α is usually chosen to be in the interval $[0.98, 1)$ [24]. Equation 2-2 is the discrete approximation of the model proposed by Flanagan [40]. Flanagan approximated the lips as a resistance in parallel with an inductance. This approximation models a piston in an infinite plane baffle.

If we assume that the glottal flow $u_g[n]$ and the vocal tract are independent, then the glottal flow derivative $\dot{u}_g[n]$ is the convolution of the glottal flow with the impulse response of the lips [4], i.e.,

$$\dot{u}_g[n] = r[n] * u_g[n]. \tag{2-3}$$

We refer to the glottal flow derivative for one pitch period as the *glottal pulse derivative*. In Section 2-4, it is shown how the transfer function of the lips can change position and can be placed in front of the glottis. In reality, the glottal flow and the glottal flow derivative of real speech signals have much more complicated structures than those of Equations 2-1 and 2-3. Six very important reasons for these complicated structures are given in the sequel.

Aspirated voicing: When a small fraction of the glottis remains slightly opened over the pitch period, a non-linear creation of turbulence in the glottal flow is present, named aspirated voicing [4]. A simplified model of aspirated voicing is additive random noise to the glottal flow signal [13].

Pitch jitter: The glottal flow for a particular vowel is quasi-periodic, which means that the pitch slightly varies in successive pitch periods. This phenomenon is called pitch jitter [4].

Diplophonia and vocal fry: In Section 2-1, we saw that the glottal flow in one pitch period may have more than one glottal pulse due to diplophonia or vocal fry. In these cases the glottal flow for one pitch period is given by

$$g[n] = g_{\text{pri}}[n] + \rho g_{\text{sec}}[n-d], \tag{2-4}$$

where $g_{\text{pri}}[n]$ is the primary glottal pulse, $g_{\text{sec}}[n-d]$ is a weaker time-shifted secondary glottal pulse, ρ is an attenuation factor and d is the delay of the secondary glottal pulse [4].

Amplitude shimmer: The amplitude of the glottal flow may vary in different pitch periods. This phenomenon is known as amplitude shimmer [4]. The pitch jitter and the amplitude shimmer are very important factors in the naturalness of speech synthesis systems [4].

Source-vocal tract non-linear interaction: The source-vocal tract independence assumption is not true in general. In voiced speech, the non-linear interaction is stronger during the open phase of the glottal flow and causes a small change of the first formant position of the vocal tract and some disturbances to the glottal flow signal [4,13]. Unlike the open phase interval, in the closed phase interval the interaction becomes very small or zero. This property is utilized in the closed-phase analysis method (see Subsection 3-3-2), which estimates accurately the vocal tract filter and the glottal flow signal. Ananthapadmanabha and Fant [37] approximated the non-linear interaction by keeping the vocal tract constant and put all of this non-linear interaction on the glottal flow derivative. Their continuous time model for the glottal flow derivative including this non-linear interaction is

$$\dot{u}_g(t) = r(t) * u_g(t) + f(t)e^{-0.5tB_1(t)}cos[\int_0^t \varOmega_1(\tau)d\tau], \tag{2-5}$$

where $B_1(t)$ and $\varOmega_1(t)$ are the time-varying bandwidth and frequency respectively, of the first formant of the vocal tract. Furthermore, $f(t)$ is an amplitude modulation function controlled by the glottal area function (i.e., the changing area of the glottis over time). The left part of this equation is the *coarse structure* of the glottal flow derivative and the right part is a sinusoidal-like component called *ripple.* Considering the non-linear interaction as a part of the glottal flow derivative is very useful because we are able to consider the vocal tract as *short-time stationary.* In words, the non-stationarity of the vocal tract is assumed to be caused only from its changing shape (which is assumed constant for short-time intervals) and not from the non-linear interaction (i.e., the changing position of the first formant). We should notice that the source filter model (see Section 2-4) is based on the short-time stationarity assumption of the vocal tract.

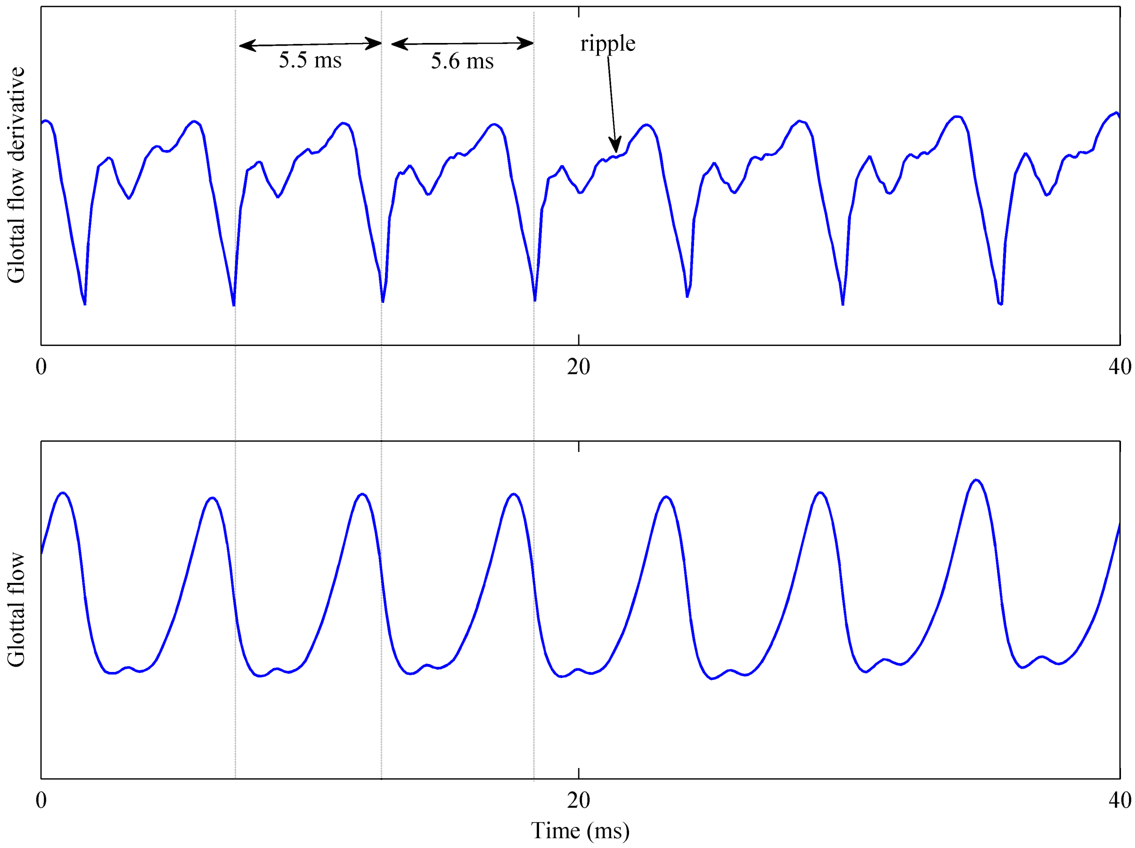


Figure 2-5: Example of an approximate real glottal flow and glottal flow derivative using IAIF method.

Source-subglottal area non-linear interaction: The subglottal reactance has been shown to delay the peak of the glottal pulse relative to that of the glottal area [41, 42]. In words, when the glottis has its maximum possible area, the glottal flow signal has not reached its maximum value yet. Moreover, in breathy phonation, additional subglottal formants appear in the speech spectrum [43].

Although there is no way of obtaining the exact glottal flow derivative of a real speech signal, we can approximate it through inverse filtering techniques [4, 44] discussed in the next chapter. An example of an approximate true glottal flow and its corresponding glottal flow derivative waveform is depicted in Figure 2-5. In this figure we can see the pitch jitter and the ripple due to the source-filter non-linear interaction.

2-2-1 Time-Domain Modeling of Glottal Pulse Derivative

In this subsection we review some well known models of the coarse structure of the glottal pulse derivative waveform. The transfer function of the glottal pulse can be modeled by a maximum-phase anti-causal all-pole transfer function with two poles outside the unit circle

$[4, 45]$

$$g[n] = (\beta^{-n}u[-n]) * (\alpha^{-n}u[-n]),\ G(z) = \frac{1}{(1-\beta z)(1-\alpha z)}, \tag{2-6}$$

where β, α are less than one. Being the output of a physical system, the speech signal is assumed to be stable, and so will be the glottal pulse. That is why the maximum-phase part of the glottal pulse is considered anti-causal, because the region of convergence (ROC) of its Z-transform has to include the unit circle [46]. Although this model approximates very well the spectral magnitude of the glottal pulse, it is not so accurate in time domain [4]. There are other more accurate models [29, 30, 47–49] in both time and frequency domains. Some of these models are more well-known for the glottal pulse derivative instead of the glottal pulse, but we can simply obtain the expression of the glottal pulse by integrating the glottal pulse derivative (i.e., by inverse filtering with the lips transfer function of Equation 2-1). The reason of preferring to model directly the glottal pulse derivative and not the glottal pulse is that the models of the former are easier to be optimized/fitted to the estimated (via inverse filtering) glottal pulse derivative [50]. For example, the time instants of the negative peaks of the glottal flow derivative are easily and accurately estimated as it is explained later in this section.

LF model: One widely used and studied model of the glottal pulse derivative is the Liljencrants-Fant (LF) model [4, 29, 47] (see Figure 2-2). Its discrete time version is given by

$$\dot{u}_g[n] = \begin{cases} 0, & \text{if } 0 \leq n < t_o \\ \frac{E_e e^{a(n-t_e)} sin[\pi(n-t_o)/t_p]}{sin[\pi(t_e-t_o)/t_p]}, & \text{if } t_o \leq n \leq t_e \\ \frac{E_e}{\beta t_a}[e^{-\beta(n-t_e)} - e^{-\beta(t_c-t_e)}], & \text{if } t_e < n < t_c = T_0 \end{cases}. \tag{2-7}$$

Since this model is periodic, an equivalent expression that we use in the present thesis is

$$\dot{u}_g[n] = \begin{cases} \frac{E_e e^{a(n-t_e)} sin[\pi n/t_p]}{sin[\pi t_e/t_p]}, & \text{if } t_o = 0 \leq n \leq t_e \\ \frac{E_e}{\beta t_a}[e^{-\beta(n-t_e)} - e^{-\beta(t_c-t_e)}], & \text{if } t_e < n < t_c \\ 0, & \text{if } t_c \leq n < T_0 \end{cases}, \tag{2-8}$$

where its parameters are defined as follows.

1. t_o: the open-phase starting time point, also called glottal opening instant (GOI), which is usually assumed to be zero in order to reduce the number of the model parameters.
2. E_e: the value of the negative peak at the instant t_e of the glottal flow derivative.
3. t_p: the zero-crossing instant of the glottal pulse derivative (i.e., the instant of the maximum glottal flow velocity). Note that, for real speech signals, the inequality $0.5t_e < t_p < t_e$ always holds[1] [51–53].
4. t_e: the return-phase starting time point [53]. Note that the second derivative of the glottal pulse at t_e is zero because its first derivative (i.e., the glottal pulse derivative) has a minimum at t_e.
5. t_c: the closed phase starting time point, also called glottal closure instant (GCI). When the LF model has to be fitted to the estimated glottal pulse derivative and the number

[1] This inequality is derived from the fact that, for real speech signals the reflection coefficient, a_m, takes values strictly in the interval $[0.5, 1]$ [51].

of parameters matters in the complexity of the fitting algorithm, sometimes a common simplification of the LF model is to assume $t_c = T_0$ [29, 51]. Moreover, we can think of this simplification as a case in which the speech signal has not an ideal closed phase. In Subsection 2-2-2 we verify that this simplification causes a non-significant change of the spectral magnitude of the glottal pulse derivative.

6. T_0: the pitch period interval which is the total duration of the open, return and closed phases.
7. t_a: characterizes the speed of the return phase and it always satisfies the inequality $t_a < t_c - t_e$ [51]. It is also called effective duration of the return phase [47] because after the point $t_e + t_a$, till the point t_c, the glottal pulse derivative is very close to zero. Note also that the return phase interval $[t_e, t_c)$ is approximately the same with a first-order low-pass filter with cut-off frequency $f_a = f_s/(2\pi t_a)$ (more about this in Subsection 2-2-2) [29].
8. β: determines how quickly the glottal pulse derivative returns to zero after time t_e and is given by the following implicit equation [29]

$$\beta t_a = 1 - e^{-\beta(t_c - t_e)}. \tag{2-9}$$

For very small t_a is given by

$$\beta = 1/t_a. \tag{2-10}$$

9. α: Roughly speaking, it determines the ratio of E_e to the maximum of the glottal flow pulse and is given by the following implicit equation

$$\frac{1}{\alpha^2 + \left(\frac{\pi}{t_p}\right)^2}\left(e^{-\alpha t_e}\frac{\pi/t_p}{sin(\pi t_e/t_p)} + \alpha - \frac{\pi}{t_p}cotg(\pi t_e/t_p)\right) = \frac{t_c - t_e}{e^{\beta(t_c - t_e)} - 1} - \frac{1}{\beta}. \tag{2-11}$$

The implicit Equations 2-9 and 2-11 guarantee area balance, resulting in a zero-net change of the glottal pulse derivative (i.e., $\sum_0^{T_0-1} \dot{u}_g[n] = 0$, which means zero mean) [29]. In reality, the glottal flow derivative is not always exactly zero-mean for each pitch period, because, sometimes during phonation, the vocal folds do not collide, which means that the glottal flow base-line is increased slightly over time [50]. As we will show in Chapter 3, the mean value of the glottal flow derivative is decreased if we apply pre-emphasis to the speech signal with a first order or a second order FIR filter which are approximations of a first order or a second order derivative, respectively. This type of pre-emphasis is important in Chapter 3, because some linear prediction unbiased estimators assume zero mean error (i.e., zero mean glottal flow derivative). In many papers (e.g. in [47, 51, 54]) the behavior in time and frequency domains of several glottal pulse derivative models is observed as a function of the following three additional parameters which are functions of the original parameters t_p, t_e, t_a, T_0.

1. The *open quotient* $O_q = t_e/T_0$ [47, 51, 54], which shows how long is the open phase with respect to one pitch period. In [54], it was undertaken a statistical analysis of several glottal pulse parameters, by taking measurements from 25 males and 20 females. It was shown that under various vocal intensities, the open-quotient values for males and females range in the intervals $[0.46, 0.91]$ and $[0.52, 0.95]$, respectively.
2. The *asymmetry coefficient* $a_m = t_p/t_e$ [51], which indicates the degree of asymmetry of the glottal flow derivative for one pitch period (e.g., for $a_m = 0.5$ we have a symmetric

sinusoidal glottal flow derivative, while for larger a_m it becomes asymmetric), usually taking values in the interval $[0.6, 0.8]$. The asymmetry coefficient of real speech signals takes values strictly in the interval $[0.5, 1]$ [51]. Note that for mathematical reasons, the LF model is capable to model the glottal flow derivative only for $a_m \geq 0.65$ [51].

3. The $R_a = t_a/T_0$ parameter [47], which indicates how long is the effective duration of the return phase over one pitch period, usually taking values which satisfy the inequality $t_a < t_c - t_e$. For the Rosenberg model, that we explain later on in this section, $R_a = 0$ because $t_a = 0$.

Here we define the new parameter $R_c = (t_c - t_e - t_a)/T_0$, which indicates how long is the non-effective return phase with respect to the whole glottal pulse. It is worth noting that in real speech signals, when the vocal intensity increases, O_q decreases (i.e., the t_e decreases), a_m increases and R_a decreases (i.e., the t_a decreases) [47, 51, 54].

Unlike the model of Equation 2-6, the LF model is mixed-phase. In particular, the open phase is modeled as an anti-causal maximum phase component, while the return phase is modeled as a causal minimum phase component [55, 56]. Moreover, it should be mentioned that in some papers (e.g. in [24, 57–59]) the GCI instants are defined as the positions of some selected large epochs of the residual of the linear prediction method. Other papers (e.g. [56]) define the GCI as the instant t_e. Maybe, the reason for all these different definitions of GCI is that many older papers (e.g. [24]) were based on the assumption that the glottal pulse derivative can be modeled accurately via the Rosenberg model [30, 51] in which $t_e = t_c - 1$.

Rosenberg model: The Rosenberg model [30] was proposed 14 years earlier than the LF model. It does not model the return phase of the glottal pulse derivative (i.e., it assumes that $t_a = 0$) and, therefore, it assumes that $t_e = t_c - 1$. It is a three-parameter model and its discrete-time version is given by

$$\dot{u}_g[n] = \begin{cases} \frac{\pi A}{2t_p} sin(\pi \frac{n}{t_p}), & \text{if } t_o = 0 \leq n < t_p \\ -\frac{\pi A}{2(t_e - t_p)} sin(\frac{\pi}{2} \frac{n - t_p}{t_e - t_p}), & \text{if } t_p \leq n \leq t_e = t_c - 1 \\ 0, & \text{if } t_c = t_e + 1 \leq n < T_0 \end{cases}, \tag{2-12}$$

where A is the amplitude of the glottal pulse at the instant t_p. Since the Rosenberg model does not model the return phase, it is a maximum-phase anti-causal model consisting only of the open phase. As was explained before, soft voices (i.e., speech with low vocal intensity) have long return phase and, thus, the Rosenberg model is inappropriate for such cases.

Figure 2-6 depicts the Rosenberg and the LF models in the time domain and their zero Z transforms (ZZT). ZZT is the Z transform of a finite-length sequence (in our case the glottal pulse) and it consists of a finite number of zeros and an equal number of poles placed at zero [4, 46]. The reason for this is that, a small finite number of non-zero poles can be computed with an infinite number of non-zero zeros and vice versa [4]. Therefore, they can be approximated very well with the M (where M is big) most significant non-zero zeros. In particular, when we have finite sequences, the value of M is the number of samples of these sequences. This can be shown in the example where we want to approximate one pole, $z = \alpha$, with a finite number of zeros using the geometric series formula

$$\sum_{k=0}^{\infty} \left(\alpha z^{-1}\right)^k \approx \sum_{k=0}^{M-1} \left(\alpha z^{-1}\right)^k = \frac{1 - \left(a z^{-1}\right)^M}{1 - \alpha z^{-1}} = \frac{z^M - \alpha^M}{z^{M-1}(z - \alpha)}, \tag{2-13}$$

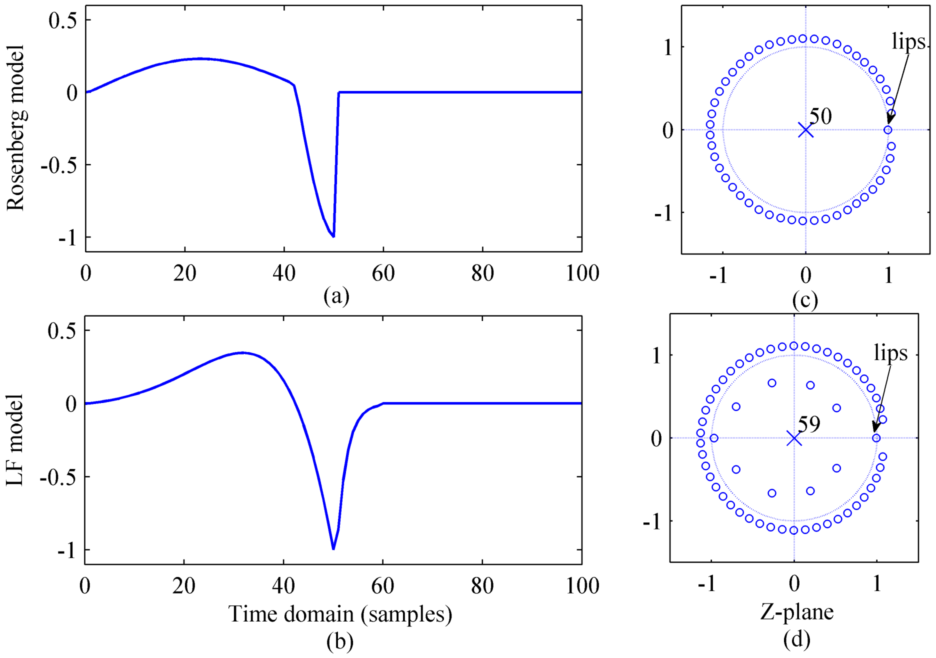


Figure 2-6: The Rosenberg model (a,c) and the LF model (b,d).

where the zeros are $z_k = \alpha e^{j(2\pi k/M)}, k = 0, ..., M-1$ and the approximated pole $z = \alpha$ is canceled from the zero $z_0 = a$. If $M \to \infty$, the sum is equivalent to $1/\left(1 - \alpha z^{-1}\right)$ and the pole, canceled previously, now is revealed. Thus, for instance, if we consider the continuous-time function (which is expressed with infinite number of zeros) of the corresponding discrete-time sequence, the Laplace transform, which is the continuous version of the Z-transform, will have one non-zero pole.

The mixed-phase property of the LF model and the maximum-phase property of the Rosenberg model are clear from Figure 2-6. The maximum-phase part is the set of zeros outside the unit circle approximating two conjugate poles outside the unit circle close to the zero frequency, while the minimum-phase part is the set of zeros inside the unit circle approximating again one or two poles. Note also that the Rosenberg model does not have a minimum-phase part because it does not have the return-phase. In Figure 2-6 (d) we see that the number of zeros of the minimum-phase part is smaller than the zeros of the maximum-phase part. This is because the return phase of the LF model has only a few samples, while its open phase has much more.

Lumped-element model: Unlike the aforementioned models, there are other more complicated mechanical models of the glottal flow, such as the lumped-element model of Story and Titze [31, 60]. This model is capable of emulating the physiological vocal folds kinematics using three masses which are coupled to one another through stiffness and damping elements (see Figure 2-7). Some of the input parameters of this model are: a) the activation levels of the cricothyroid muscle and the throarytenoid muscle, b) the resting length and thickness of the vocal folds, c) the prephonatory glottal half-width at the inferior and superior edges of the vocal folds, and d) the respiratory pressure applied at the entrance of the trachea. A limitation of this model is its low-dimensionality, due to the representation of the vocal folds

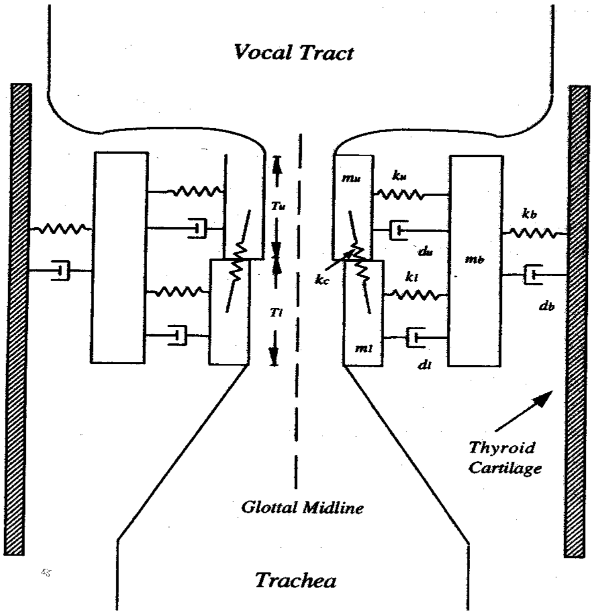


Figure 2-7: Lumped-element vocal fold model.

SOURCE: [31].

with a small number of bar masses. As a consequence, there is no anterior-posterior glottal variation in the opening or closing phases, and this leads to an abrupt closure and opening. This of course results in discontinuities in the glottal flow derivative as illustrated in Figure 2-8. Note that the glottal flow of Figure 2-8 has been generated using the software LeTalker [61] considering supraglottal and subglottal interactions, and the glottal flow derivative is computed using Equation 2-3.

Now we give an example to justify the previous discussion about the different definitions of GCI. The different shapes of the glottal flows and glottal flow derivatives, modeled by the LF

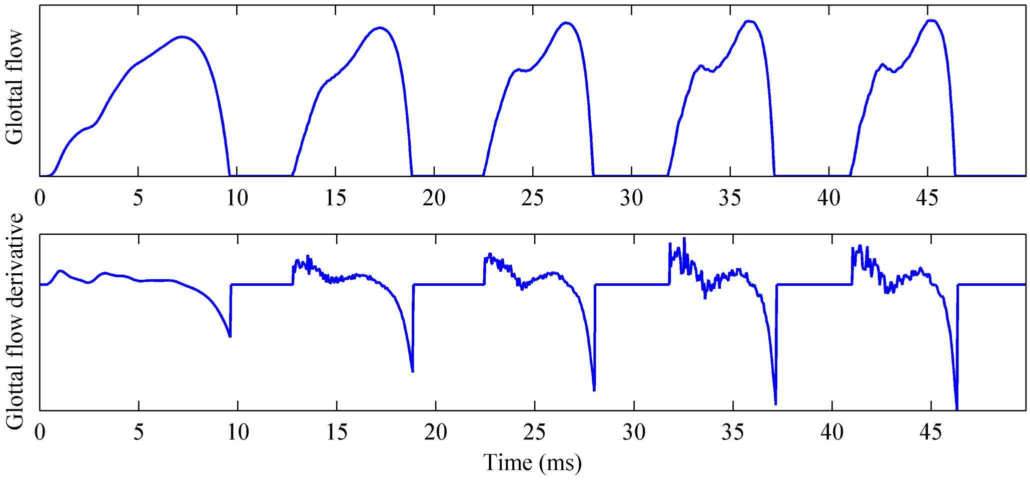


Figure 2-8: The glottal flow and the glottal flow derivative generated via the LeTalker software which is based on the lumped-element vocal fold model.

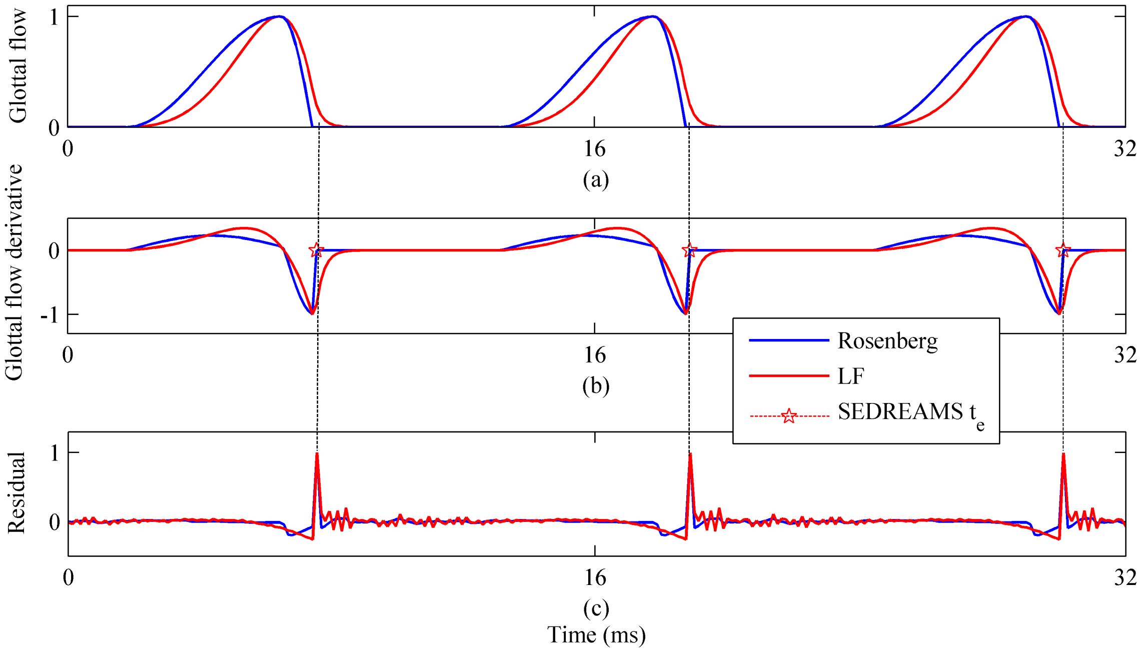


Figure 2-9: Glottal flows (a), glottal flow derivatives (b), and residuals obtained through linear prediction (c), using the LF model and the Rosenberg model.

and Rosenberg models, are depicted in Figure 2-9 (a,b). Note that in case of the Rosenberg model, the glottal flow derivative becomes zero at the instants t_c while the corresponding glottal flow becomes zero at one sample before, i.e., at $t_c - 1$ (the same result is shown in [24]). Furthermore, the instant $t_c - 1$ of the Rosenberg model is equal to the t_e instant of the LF model.

GCI estimation: There are two different definitions of GCI in the literature. According to the first one, GCIs are the t_e instants which are one sample before the instants of the large epochs of the residual of linear prediction [14, 24, 57–59]. According to the second definition, which is used in the present thesis, GCIs are the t_c instants, i.e., the instants where the closed phases of the glottal flow start [4, 13]. To the author's opinion, the reason of the prevalence of the first definition in many recent papers [56–59] is that some previous papers [14, 24] were based mostly on the assumption of the Rosenberg model (i.e., $t_c = t_e + 1$) and assumed reasonably that the large epochs of the residual of a synthesized signal, using the Rosenberg model, occur at the t_c instants (see Figure 2-9) (b,c). However, we know that the closed phase for soft-voices does not start at t_e but at t_c occurring some samples after t_e [13, 29, 37, 47, 48].

There are many methods estimating the t_e instants through the large epochs of the residual [57–59]. The estimation of the t_e instants can be used in many applications such as: in glottal flow derivative model fitting; in speaker localization [15]; in determining the closed phase interval for the closed-phase analysis methods [24].

The latter application may be problematic for soft voices, i.e., when $t_c > t_e$. Of course, it is much easier to estimate the GCI positions using the large epochs of the residual than trying to estimate directly the t_c instant. Thus, one may say that since the true glottal closure instant t_c is a few samples after the epoch of the residual, we can just add a small number of

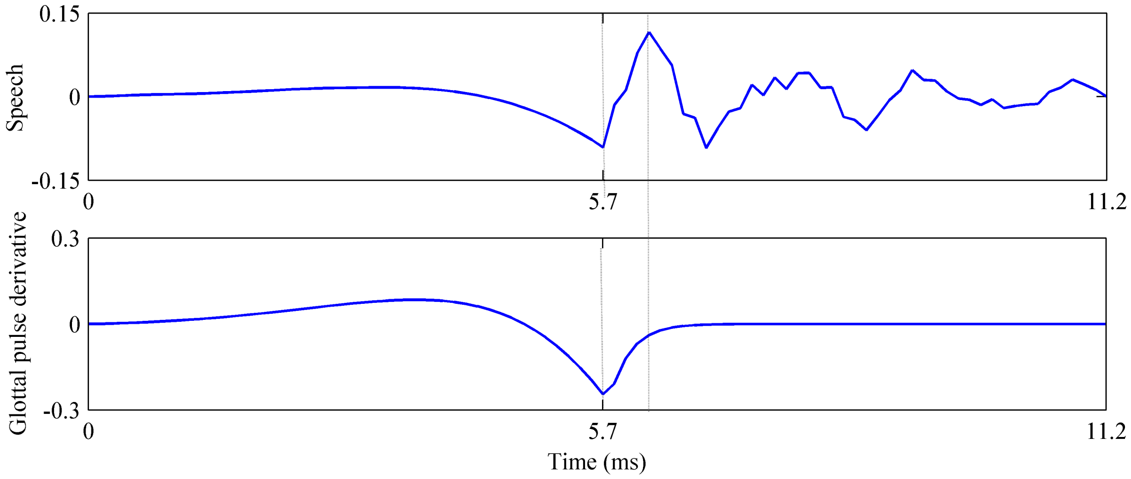


Figure 2-10: Proposed method of estimating the instant t_c.

samples to the estimated position of a large epoch and obtain an estimate of the t_c instant. But the question that arises here, is how many samples? In the present thesis we propose a method of estimating this number of samples (we are not aware of other similar methods proposed previously). The proposed method is based on the observation that the speech signal close to the instant t_c and closer to the instant $t_e + t_a$ (i.e., slightly after the instant t_e) has its maximum value (see Figure 2-10). So, to our opinion, one idea of estimating the starting point (which is very close to the t_c instant) of the interval that is used for closed-phase analysis, is to find first the instant t_e with one of the existing methods in the literature [57–59] and then search for the global maximum value of the speech signal in the next few samples, and set as the t_c instant this point. We leave for future investigation a more detailed exploration of the robustness of this method in real speech signals.

Another method [13] for determining the closed phase interval tries to estimate the time interval where the vocal tract filter remains constant/stationary. This method applies sliding covariance analysis with one sample shift and defines as the closed phase interval the interval where the first formant of the estimated vocal tract remains constant/stationary.

Figure 2-9 (b) shows the results of a small experiment for estimating the t_e instants (denoted with red stars) using he SEDREAMS algorithm [24]. For this purpose, a synthetic speech signal was generated by convolving the LF model for multiple pitch periods with an all-pole filter. As we can see, the estimated t_e positions coincide on the large epochs of the residual.

2-2-2 Spectral Characteristics of the Glottal Pulse Derivative

The spectral magnitude of the glottal pulse derivative consists of two main parts: the *glottal formant* which is at the low frequencies, and the *spectral tilt* [29] which is after the glottal formant and covers the longest part of the spectral magnitude. Note that, the glottal formant is the maximum-phase component of the glottal pulse derivative, i.e., the two conjugate poles outside the unit circle close to the real axis of the Z-plane. While the LF model can describe changes in the spectral tilt [51], the Rosenberg model cannot due to its inability to encompass the return phase. Thus, the Rosenberg model has a drawback since modeling the spectral

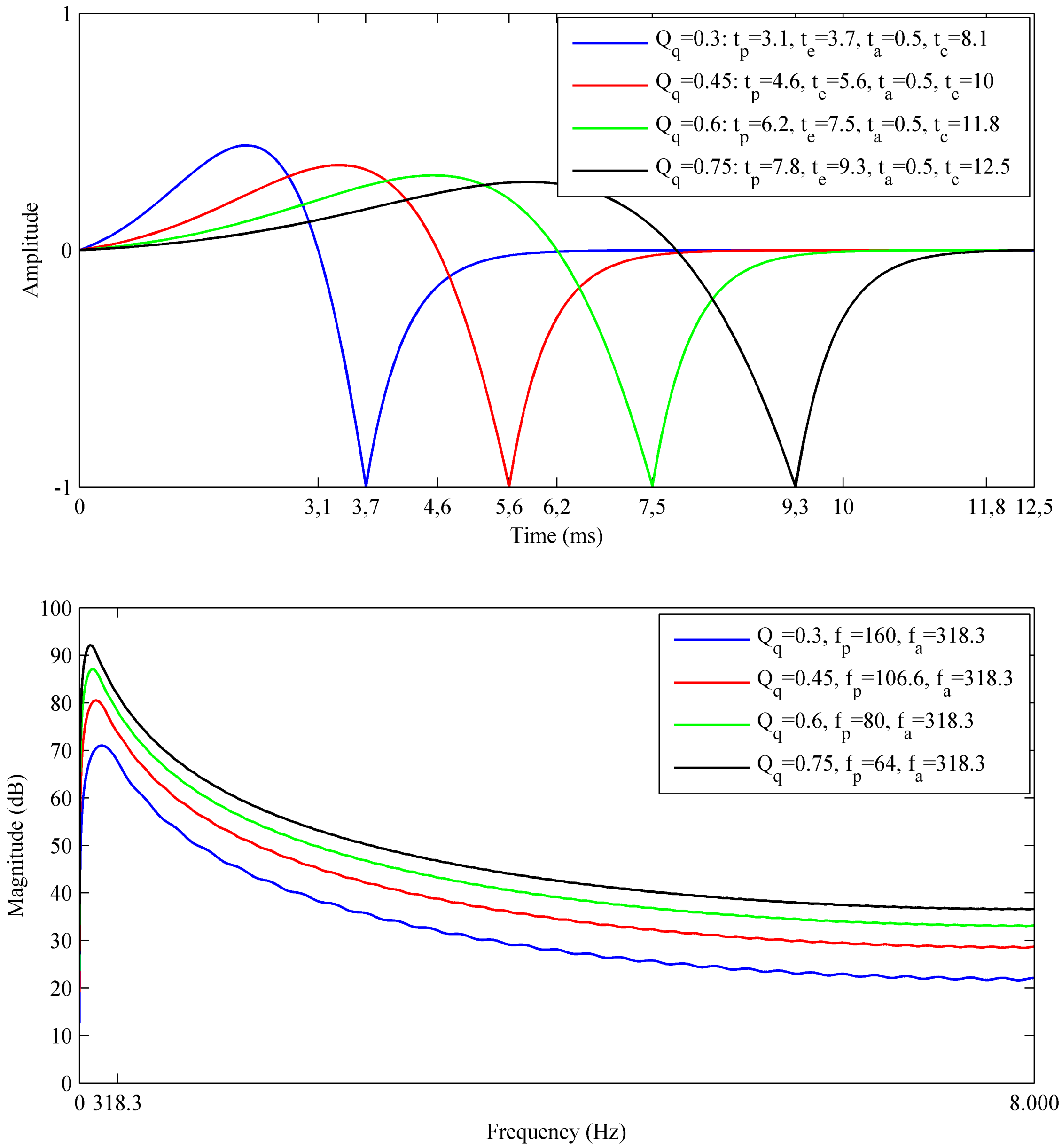


Figure 2-11: O_q effect on Magnitude.

tilt changes may be very important as in prosodic stress voice perception [62]. The changing values of the parameters Q_q, a_m and R_a are related to particular changes in the spectral magnitude of the glottal pulse derivative [29, 47, 51]. In this section we summarize these changes, and additionally, we explore the changes that are caused by the parameter R_c.

In [63], it was shown that the glottal formant occurs in frequency $f_p = f_s/2t_p$, where f_s is the sampling frequency. We should mention at this point that, since $t_p < T_0$, f_p should be strictly greater than $F_0/2$, where F_0 is the pitch. Moreover, since the open quotient usually takes values in the interval $[0.46, 0.95]$ (see Subsection 2-2-1), it can be derived[2] that f_p usually

[2] For this derivation use the formula of the open quotient and the inequality $0.5t_e < t_p < t_e$.

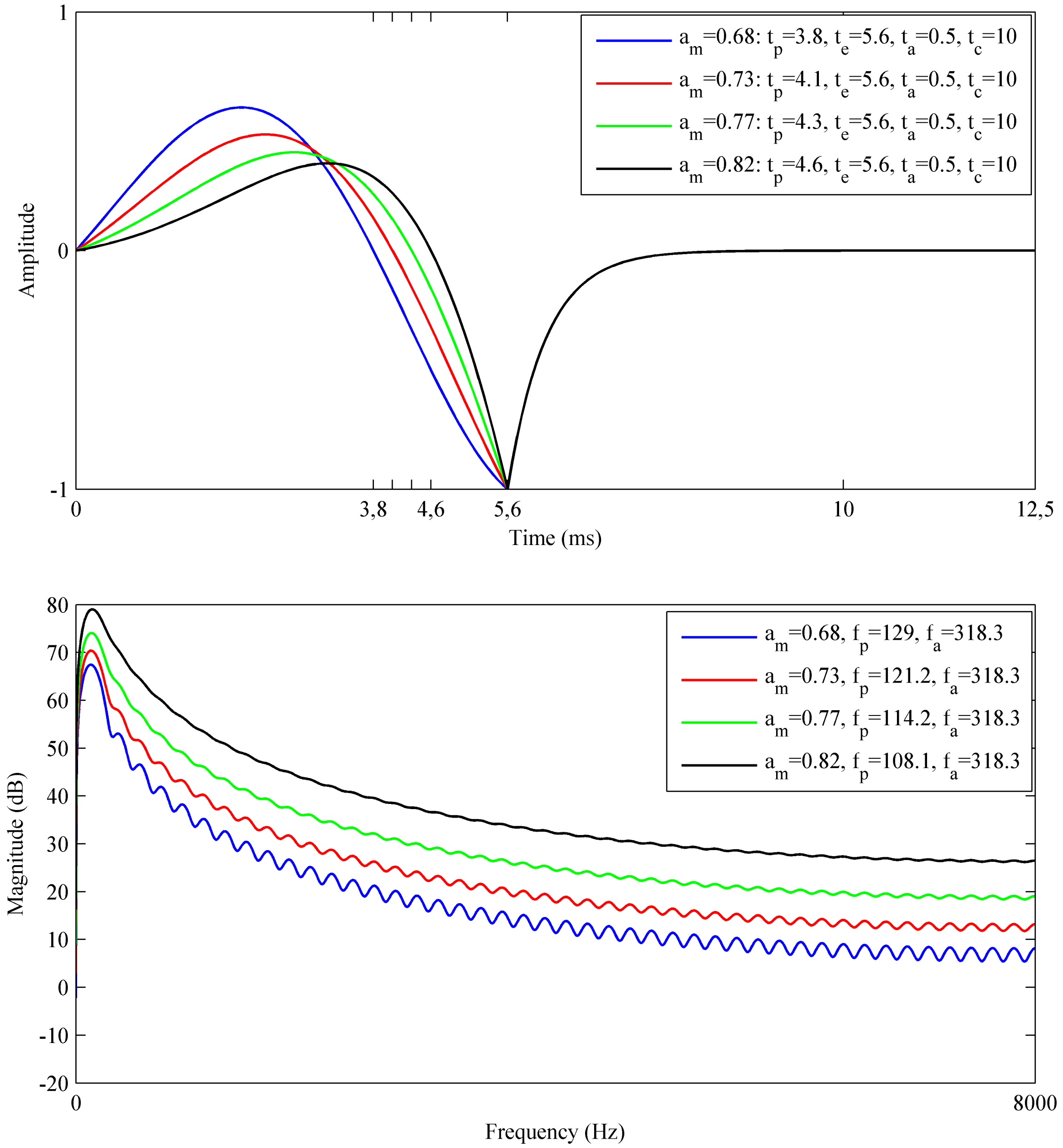


Figure 2-12: a_m effect on Magnitude.

does not take greater values than $F_0/0.46$. Thirteen years before the development of the LF model, Wakita [64] assumed a -12dB/octave[3] spectral magnitude roll-off of the glottal pulse and a -6dB/octave spectral magnitude roll-off of the glottal pulse derivative, because the lips can be considered as a high-pass filter with a +6dB/octave magnitude increment. The -6dB/octave magnitude roll-off is achieved approximately, when there is no return phase in the glottal pulse derivative (as was explained before this happens when the vocal intensity is high and there is an abrupt change after the instant t_e). On the other hand, the glottal pulse

[3] Octave bands are frequency bands in which the highest band frequency is twice the lowest band frequency. The number of octaves between two frequencies f_1 and $f_2 > f_1$ is $\log_2(f_2/f_1)$.

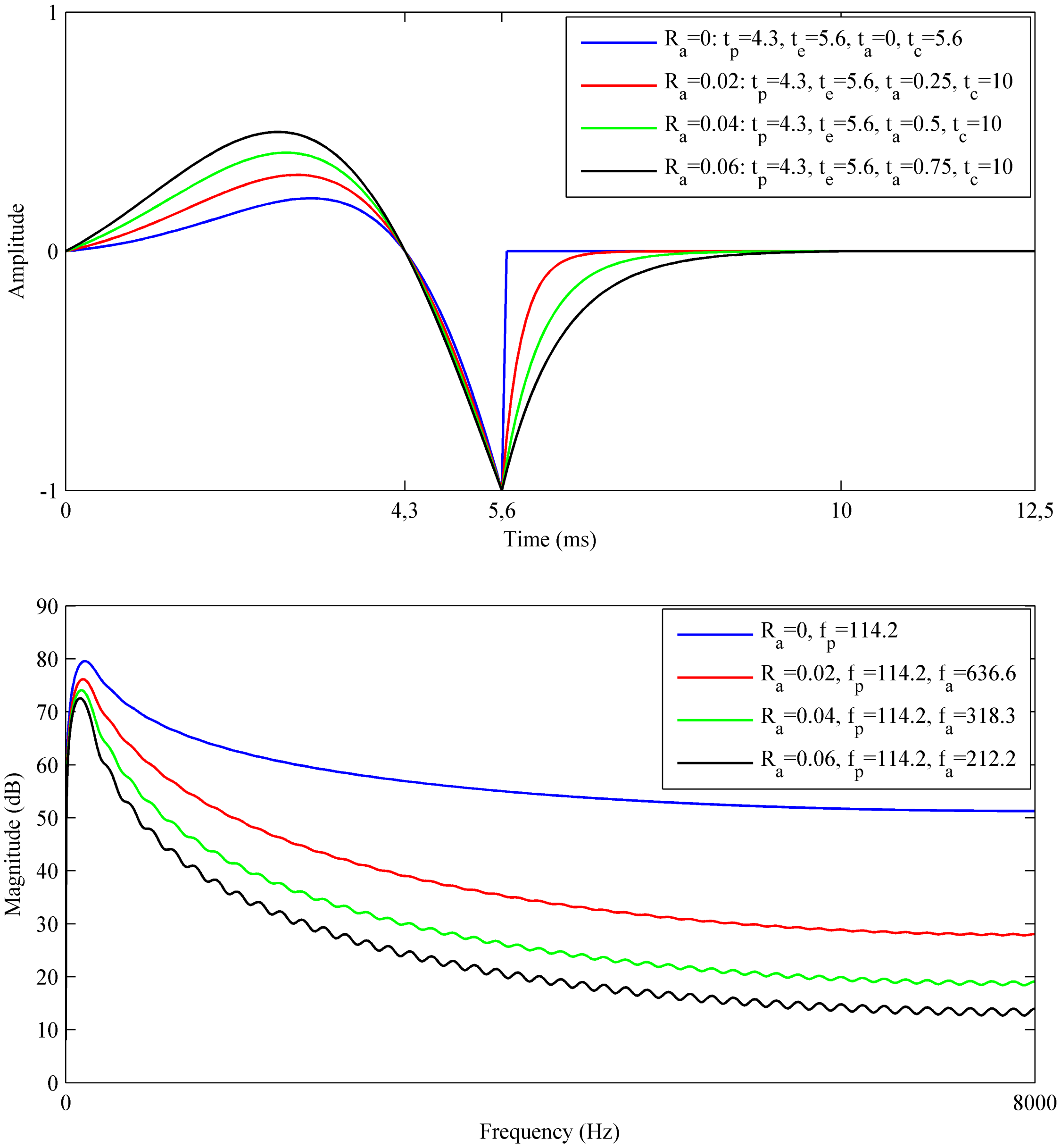


Figure 2-13: R_a effect on Magnitude.

derivative of soft-voices has a return phase and, according to the LF model there is an extra -6dB roll-off to the spectral tilt of the magnitude [29, 47]. This is because the return phase component is an exponential function and can be approximated very well with a low-pass filter with cut-off frequency $f_a = f_s/(2\pi t_a)$ [47], where t_a is the effective duration of the closed-phase interval. Therefore, the extra -6dB/octave roll-off will occur for $f > f_a$.

Now, we summarize some of the most important results of [29, 47, 51] about the effects of the changing parameters of the LF model in the spectral magnitude of the glottal pulse derivative. As we can see in Figure 2-11, an increase of O_q, keeping constant the parameter a_m, means a decrease of the glottal formant frequency and a decrease of the spectral magnitude,

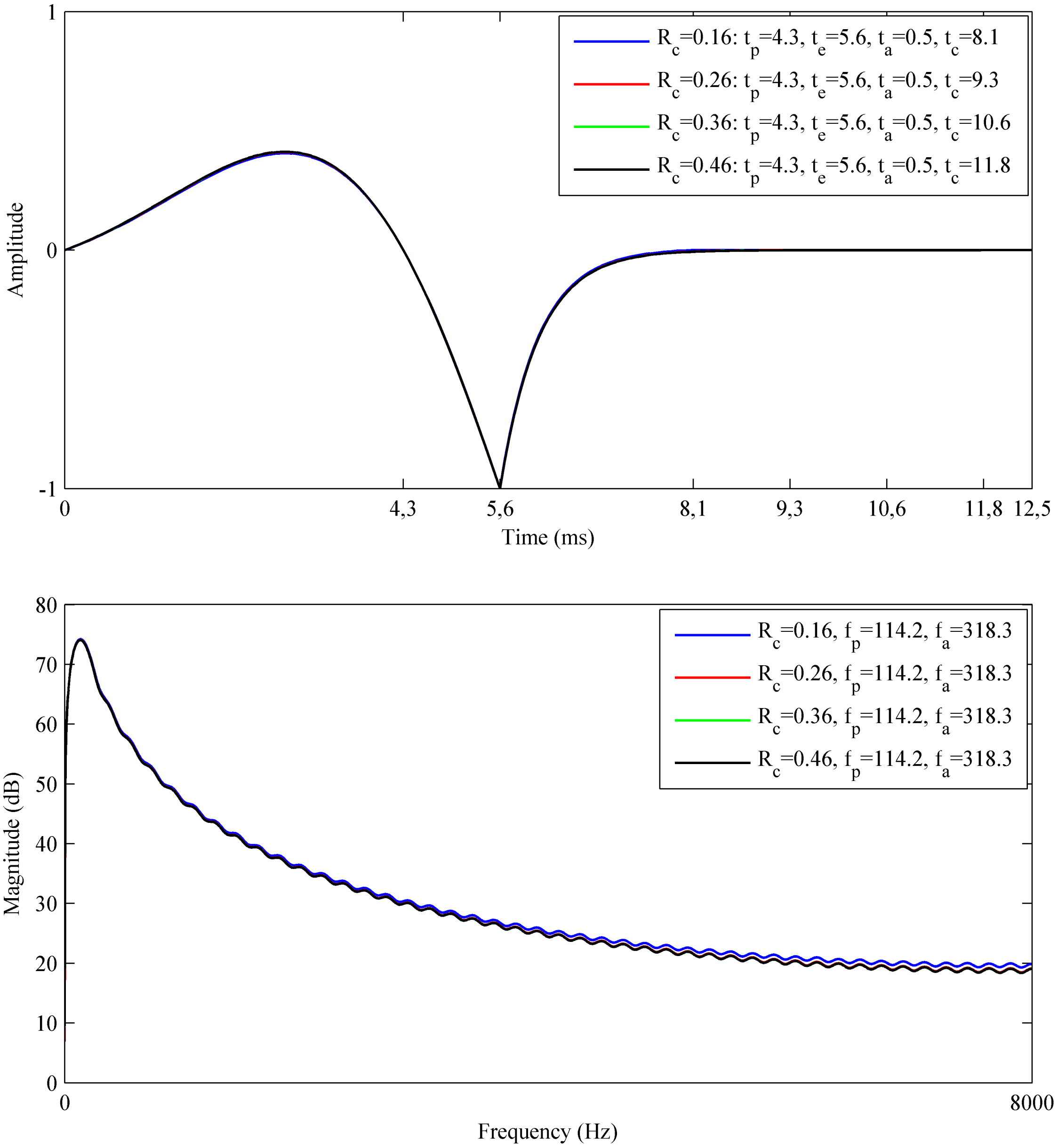


Figure 2-14: Rc effect on Magnitude.

almost equivalently for all frequencies, which means that the spectral magnitude roll-off is not effected. A similar situation happens with the parameter a_m (see Figure 2-12) keeping constant the parameter O_q. On the other hand, the only parameter that effects the spectral magnitude roll-off is R_a, because of the changing t_a. As it is shown in Figure 2-13, an increased R_a means that the additional -6dB/octave starts in lower frequencies. Finally, in Figure 2-14 we see that the parameter R_c effects negligibly the spectral magnitude of the glottal pulse derivative. This explains why the commonly used assumption of $T_0 = t_c$ in the LF model works correctly.

2-3 Modeling of Vocal Tract

As was explained in Section 2-1, the vocal tract changes shape for different types of vowels, consonants etc. According to the discrete concatenated tube model, the shape of the vocal tract can be characterized by its discrete *area function* [34], which is given by

$$f(L_n) = A_n, \text{ for } n = 1, 2, ...L, \tag{2-14}$$

where A_n are the cross-sectional areas of the L concatenated tubes and L_n their distances from the glottis. In reality, not only the cross-sectional areas change, but also their distances from the glottis (i.e., the lengths of the concatenated tubes) and, consequently, the total length of the vocal tract [34]. The variation of the distance between two neighboring points in the vocal tract is typically a millimeter or less, and if we sum all these variations, the total vocal tract length can vary one centimeter or more [34]. The average vocal tract length of a male speaker is 17 cm while for females and kids is less [4].

According to the Kelly-Lochbaum model [32], also called the wave reflection analog of the vocal tract, the vocal tract is a forward-backward wave system. It is modeled with a uniaxial set of equally long tubes with different cross-sectional areas which are related to a set of *reflection coefficients* as

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}, \tag{2-15}$$

where A_k is the cross-sectional area of the k-th tube. Note that if either A_{k+1} is greater than A_k, or A_k is greater than A_{k+1}, the reflection coefficients satisfy the inequality $|r_k| \leq 1$. This simplistic model is based on the lossless-tube assumption and it does not encounter losses in the vocal tract. Thus, the Kelly-Lochbaum model assumes no energy loss inside the vocal tract but only at the lips and the glottis [4]. Based on the Kelly-Lochbaum model, it can be proved that the vocal tract with closed velum, is approximated by an all-pole minimum-phase linear filter [4], which is given by

$$V(z) = \frac{A}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}, \tag{2-16}$$

where A is its gain, p is its order and α_k are its coefficients, found from the reflection coefficients with the step-up Levinson-Durbin recursion [65] (see Algorithm 1). Therefore, we can obtain the a_k coefficients if we know the cross-sectional areas A_k. It is worth noting that the impulse response of the vocal tract, $v[n]$, is a real sequence and, therefore, the poles will be in conjugate pairs (except of those which are on the real axis) [4]. Thus, its transfer function is given by

$$V(z) = \frac{A}{\prod_{k=1}^{p/2}(1 - d_k z^{-1})(1 - d_k^* z^{-1})}, \tag{2-17}$$

where (d_k, d_k^*) are the conjugate pole-pairs. By using the fact that $|r_k| \leq 1$, it can be proved that all these pole pairs are always inside the unit circle [4]. Note that the number of poles, p, is two times the number of the concatenated tubes or the number of formants. Since the impulse response of the vocal tract is a real sequence, the frequencies and the bandwidths of the formants can be determined by the poles with positive frequencies $z_i = r_i e^{jw_i}$, where the frequency of each formant is $(w_i f_s)/(2\pi)$ and its bandwidth is $-(\ln(r_i) f_s)/\pi$ with f_s the sampling frequency. In Chapter 3, we will see the most popular methods of estimating the coefficients α_k.

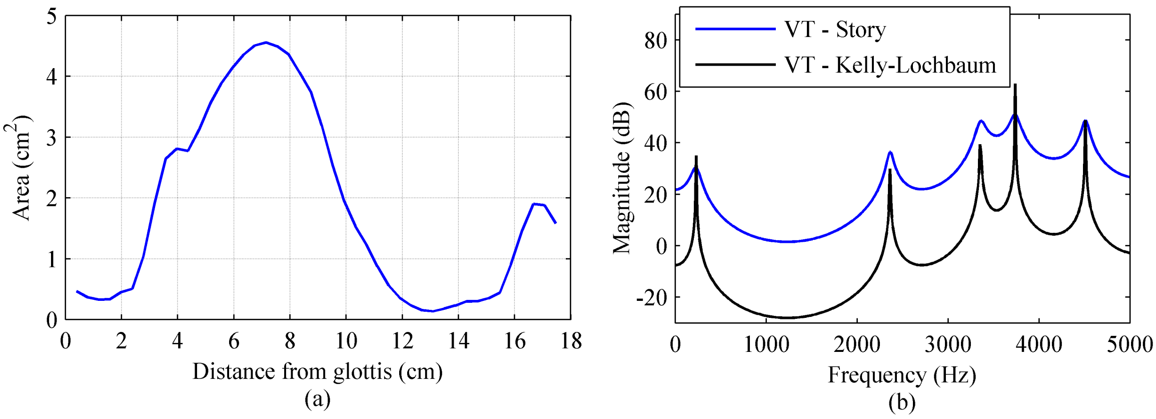


Figure 2-15: The vocal tract area for the vowel /i/ (a) is used by the Story model and the Kelly-Lochabaum model for the synthesis of the vocal tract transfer function (b).

```
Initialization: $\alpha_0(0) = 1$;
for $j = 0, 1, ..., p - 1$ do
    for $i = 1, 2, ..., j$ do
        $\alpha_{j+1}(i) = \alpha_j(i) + r_{j+1}\alpha_j(j - i + 1)$;
    end
    $\alpha_{j+1}(j + 1) = r_{j+1}$;
end
```

Algorithm 1: Levinson-Durbin step-up recursion.

We should mention here that if we try to synthesize the vocal tract transfer function[4] using the Kelly-Lochbaum model, the formants of the vocal tract magnitude will have zero bandwidth. This happens because, the Kelly-Lochbaum model assumes zero losses in the vocal tract [4]. A more accurate and complicated model of the vocal tract, proposed by Story [33], includes many possible losses of the vocal tract such as: losses of the vibrating walls, viscous fluid losses, heat conduction losses and kinetic pressure drop [33]. Therefore, the magnitude of the vocal tract will not have zero-bandwidths any more. Note that both models have identical formant locations. Figure 2-15 demonstrates an example of an area function of the vocal tract for the vowel /i/ of a male speaker acquired via magnetic resonance imaging (MRI) [34]. It also shows the spectral magnitudes of its transfer functions computed according to the aforementioned two models. The spectral magnitudes and the area function of Figure 2-15 were generated via LeTalker software written by Professor Story [61].

From Section 2-1 we know that, when the velum is opened, the vocal tract consists of zeros/anti-resonances as well. Although the poles are reasonably assumed to be always inside the unit circle, some of the zeros may lay outside the unit circle (e.g. for most of consonants such as fricatives and nasals and for impulsive speech) [4, 66]. In such cases, the vocal tract is a pole-zero mixed-phase filter given by

$$V(z) = A\frac{\sum_{k=0}^{q}\beta_k z^{-k}}{1 - \sum_{k=1}^{p}\alpha_k z^{-k}}, \tag{2-18}$$

[4] Note that in the present thesis we do not consider the lips as a part of the vocal tract (see Section 2-1).

where q is the number of zeros.

It should be mentioned that in many applications, such as linear prediction coding, the all-pole minimum phase assumption for the vocal tract is convenient and works well in terms of the perceptual quality, even when the velum is opened [4]. This is because the spectral peaks (i.e., the resonances) are more important perceptually than the spectral valleys (i.e., anti-resonances) in the spectral magnitude of the vocal tract [35].

2-4 Modeling of Speech with the Source Filter Model

An approximate discrete realization of the speech production can be achieved via the source filter model (SFM) [3, 4]. SFM assumes that the source, $u[n]$, and the vocal tract, $V(z)$, are independent. It also assumes that the vocal tract remains constant/stationary for short time intervals, which are called *frames* and are usually 20-40 ms. Therefore, the SFM models the speech frames and not the entire speech signal. Figure 2-16 shows an example of how a voiced speech signal over one pitch period is generated according to SFM in time and in frequency domains. As can be seen, due to the independence of SFM, the speech signal is generated through the convolution of the glottal pulse, the lips and the vocal tract. Also note, for the simple example of Figure 2-16 we used the LF model for constructing the glottal pulse derivative excluding the non-coarse structure of the glottal pulse derivative such as possible aspiration or the ripple component.

In this section, we consider four different versions of SFM. Each version differs in the definition of the source and the filter. In all versions, the source signal is a function of another signal, called *excitation*, which can be either a periodic impulse train for voiced speech or a zero-mean white Gaussian noise (WGN) with standard deviation 1 for unvoiced speech. The filter for all SFM versions is denoted by $H(z)$ and $h[n]$ in frequency and time domains, respectively. SFM assumes that only one of the two sources (i.e., voiced or unvoiced) can be selected in each time period. However, this simplification is not accurate when we have, for example, voiced plosives, voiced fricatives and aspirated voicing. In such cases, the source signal may be a linear or a non-linear combination of the two kinds of sources [4]. The three first SFM versions are used in all-pole speech analysis methods reviewed in Chapter 3. The fourth SFM version is for pole-zero analysis methods [4, 5, 66–68], which are not discussed in the present thesis.

2-4-1 SFM Version 1

In the first version of SFM (Figure 2-17), the source is defined as

$$u[n] = \begin{cases} u_g[n], & \text{if voiced} \\ w[n] \text{ (WGN)} , & \text{if unvoiced} \end{cases} . \tag{2-19}$$

As was already discussed in Section 2-2, the transfer function of the lips can change position and can be placed after the source, because we have a cascade connection of linear systems [46]. Therefore, we may consider as source the output of the newly positioned transfer function of the lips. If the source is voiced, $\dot{u}_g[n] = u_g[n] * r[n]$ (see Section 2-2), while, when the source is WGN, $\dot{w}[n] = w[n] * r[n]$ which is a moving average (MA) process [65]. Therefore, the final

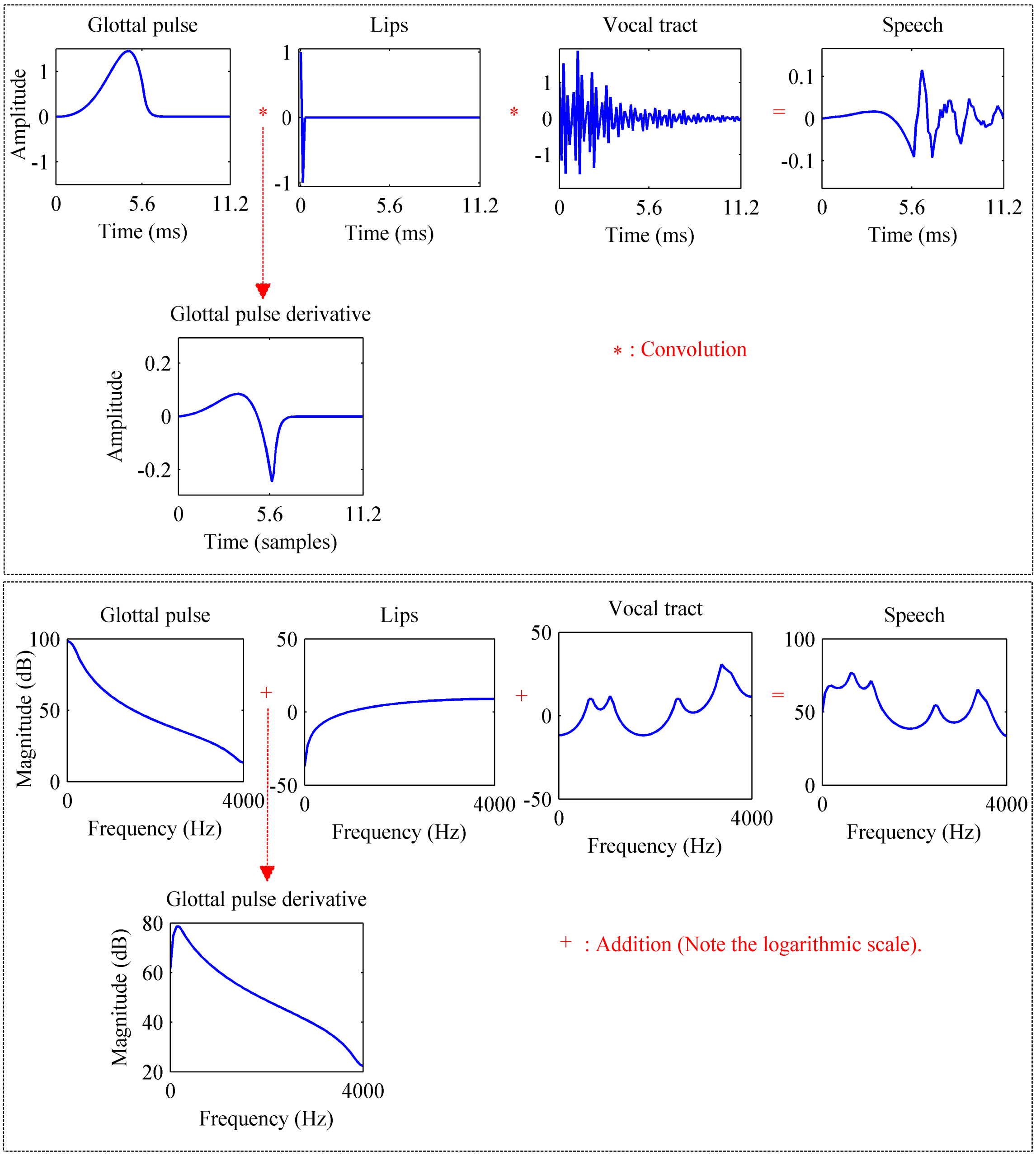


Figure 2-16: Generation of a voiced speech signal for one pitch period according to SFM in time (top box) and in frequency (bottom box).

source is the outcome of the lips transfer function and is given by

$$\dot{u}[n] = \begin{cases} \dot{u}_g[n], & \text{if voiced} \\ \dot{w}[n], & \text{if unvoiced} \end{cases}. \tag{2-20}$$

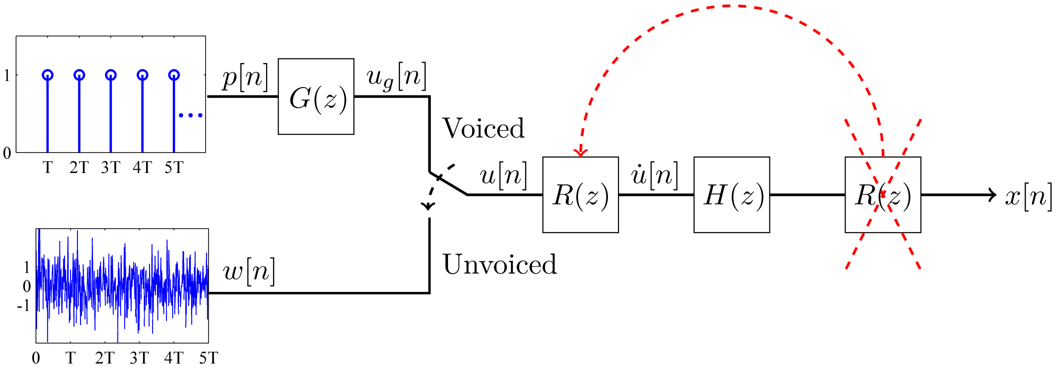


Figure 2-17: Source Filter Model Version 1.

The filter in this SFM version is equal to the all-pole minimum-phase vocal tract filter of Equation 2-16, i.e., $H(z) = V(z)$ and, thus, the speech signal can be decomposed as

$$X(z) = \dot{U}(z)H(z) = \begin{cases} \dot{U}_g(z)H(z), & \text{if voiced} \\ \dot{W}(z)H(z), & \text{if unvoiced} \end{cases}. \tag{2-21}$$

Although $H(z) = V(z)$ is considered to be minimum-phase all-pole filter, the speech model $X(z)$ is pole-zero due to the transfer function $R(z)$ and, also, mixed-phase in the case of voiced speech due to the poles of $G(z)$ which are outside the unit circle.

2-4-2 SFM Version 2

In this SFM version (Figure 2-18) the filter is

$$H(z) = \begin{cases} G(z)V(z), & \text{if voiced} \\ V(z), & \text{if unvoiced} \end{cases}. \tag{2-22}$$

This is a mixed-phase all-pole filter for voiced speech and a minimum-phase all-pole filter for unvoiced speech. Now, the transfer function of the lips is moved after the signals $p[n]$ and $w[n]$. Therefore, the source is

$$\dot{u}[n] = \begin{cases} p[n] * r[n] = \dot{p}[n], & \text{if voiced} \\ w[n] * r[n] = \dot{w}[n], & \text{if unvoiced} \end{cases}, \tag{2-23}$$

where $r[n] = 1 - \alpha\delta[n-1]$. Herein, in voiced case the source signal becomes a new impulse train signal having two impulses per pitch period instead of one. As in the SFM version 1, the unvoiced source is a MA process.

2-4-3 SFM Version 3

The third SFM version (Figure 2-19) combines the transfer functions $G(z), R(z)$ and $V(z)$ and forms the filter $H(z)$. It also "gets rid" of the zero, from the transfer function $R(z)$, by approximating it with a finite number of poles. As was explained in Subsection 2-2-1, one

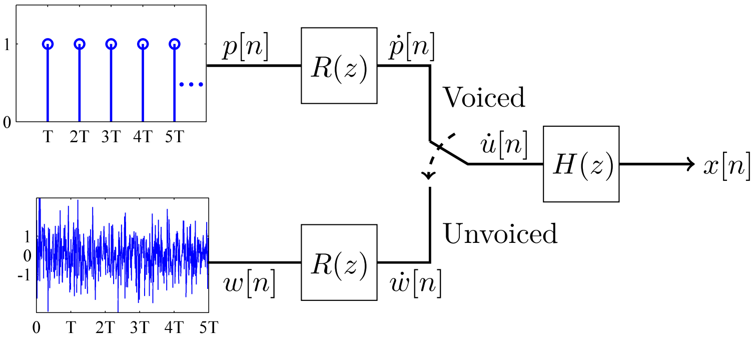


Figure 2-18: Source Filter Model Version 2.

zero inside the unit circle can be written as an all-pole transfer function of infinite many poles inside the unit circle,

$$R(z) = 1 - az^{-1} = \frac{1}{\sum_{k=0}^{\infty} \lambda^k z^{-k}}. \tag{2-24}$$

Thus, the transfer function $R(z)$ can be approximated by keeping only the M poles that contribute the most, i.e.,

$$R(z) \approx \frac{1}{\sum_{k=0}^{M} \lambda^k z^{-k}}. \tag{2-25}$$

The value of M is usually selected to be 4 [4]. Therefore, the source in this SFM version is

$$u[n] = \begin{cases} p[n], & \text{if voiced} \\ w[n], & \text{if unvoiced} \end{cases}, \tag{2-26}$$

and the filter is

$$H(z) = \frac{A}{1 - \sum_{k=1}^{p+2+M} a_k z^{-k}}, \tag{2-27}$$

where $p + 2 + M$ is the total number of the poles of $R(z), G(z)$ and $V(z)$.

The total number of poles in this SFM version is usually selected to be 16 for voiced speech [4], because the lips contribute with 4 poles inside the unit circle, the glottal flow contributes with 2 poles outside the unit circle and the vocal tract with 10 poles inside the unit circle. Therefore, in case of voiced speech, the filter is all-pole mixed-phase. It is worth noting that in unvoiced speech, in which the vocal tract might have some zeros as well (see Section 2-3), if we apply the same strategy of approximating the zeros with a finite number of poles, the number of total poles has to be increased significantly. Finally, according to SFM version 3, the speech signal can be decomposed as

$$X(z) = U(z)H(z) = \begin{cases} P(z)H(z), & \text{if voiced} \\ W(z)H(z), & \text{if unvoiced} \end{cases}. \tag{2-28}$$

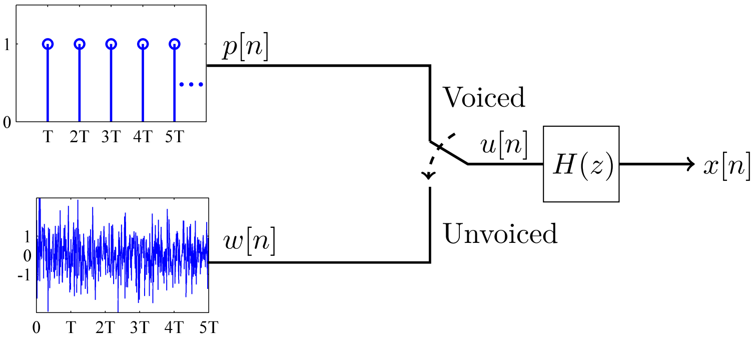


Figure 2-19: Versions 3 and 4 of the Source Filter Model.

2-4-4 SFM Version 4

The fourth version of SFM has the same source as version 3 of SFM, and combines the transfer functions $G(z), R(z)$, $V(z)$, and forms the filter $H(z)$. The only difference with the version 3 of SFM is that the filter $H(z)$ is now pole-zero. Thus, the filter is

$$H(z) = A\frac{\sum_{k=0}^{q} b_k z^{-k}}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{2-29}$$

This SFM model is appropriate when we have unvoiced speech, aspirated voicing, nasalized speech or impulsive speech, and can be used in pole-zero analysis methods. We can also use an all-pole model in aforementioned cases of speech, but, as was explained in Subsection 2-4-3, the necessary number of poles for the approximation of the zeros is high. Although the pole-zero analysis methods that use the SFM version 4 are not linear (i.e., the complexity is high), they can reduce significantly the number of transmitted parameters in speech coding applications for certain categories of speech [4, 5].

Chapter 3

Linear Prediction Analysis

Linear prediction (LP) analysis methods, also called all-pole speech analysis methods, are based on the SFM versions 1, 2 and 3 (i.e., they assume that the filter is all-pole). Roughly speaking, there are two main categories of LP methods targeting at different goals. The first category, referred in the literature as inverse filtering, aims to approximate accurately the true glottal flow or the glottal flow derivative and the true vocal tract filter and is based mostly on SFM version 1. This category of LP methods is useful in many applications such as speech synthesis [7, 8], analysis of vocal emotions [10, 11], analysis of pathological voices [12] or speaker identification [13]. For instance, in Section 2-2 it was shown that, in case of vocal fry or diplophonia, the structure of the glottal flow waveform consists of two glottal pulses instead of one over one pitch period. This characteristic can be used in determining whether a speaker has this kind of speech disorder. Moreover, in speaker identification, there are several components of structure of the glottal flow derivative which are "unique" for each speaker and, therefore, they can be used as features in speaker identification.

The second category of LP analysis methods aims to find the sparsest possible residual no matter whether the corresponding estimated filter is very close or very far from the true vocal tract filter. This category of LP methods is based on SFM versions 1, 2 and 3 and is useful in applications such as LP speech coding [4], speech enhancement/dereverberation [1], epoch extraction [14] or speaker localization [15]. In speech coding applications, where both the residual and the filter are transmitted, we want to find the sparsest possible residual while keeping the filter order as low as possible. On the other hand, in speaker localization we do not care about the order of the filter, but only about obtaining a very sparse residual. Specifically, in this application a speech signal is acquired via several microphones placed in different positions. Therefore, each acquired speech signal will have a different delay. Then, each acquired speech signal is analyzed obtaining a residual. The main epochs of the residual are used for the localization purposes. The main epochs of a sparse residual are distinguished easier and can be used for the determination of the time difference of arrivals (TDOA)s between the source and the microphones. Note that if we increase the order, the complexity will increase as well. Thus, if we want a real-time speaker localization system, the order should not be extremely high. Note that in Chapter 4, we show that when a pre-emphasis or a glottal-cancellation filter is used prior to LP analysis, a very sparse residual is obtained using a low filter order.

As was explained in Chapter 2, the speech signals are assumed to be short-time stationary and, therefore, all the analysis methods that we discuss are frame-based with frame lengths of 20-40 ms. The frames can be voiced, unvoiced or mixed, but here we explore only the voiced frames and, therefore, the source signal is considered to be the glottal flow, or the glottal flow derivative, or a periodic impulse train, depending on the SFM model that we choose.

In this chapter we review the basic properties of seven LP analysis methods which are special cases of a general LP problem. In Section 3-1 we explain how the three first SFM versions (discussed in Section 2-4) can be utilized from LP analysis and we present the general LP problem. In Section 3-2, various pre-emphasis and glottal flow cancellation techniques are presented. Section 3-3 reviews seven LP methods and explores some of their most important properties.

3-1 General Linear Prediction Analysis Problem

The LP analysis methods presented in this section are based on the SFM versions 1, 2, 3. Depending on what we want to estimate in a certain application we select one of the three SFM versions. In inverse filtering applications which aim to estimate accurately the true glottal flow derivative and the true vocal tract, we use the SFM version 1 to model an N-samples speech frame signal, $x[n]$, as

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + A\dot{u}_g[n], \text{ for } n = 0, 1, ..., N-1. \tag{3-1}$$

This is the output of the all-pole filter

$$H(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{3-2}$$

with the glottal flow derivative $\dot{u}_g[n]$ as the input/source. The glottal flow is obtained by a simple integration of the glottal flow derivative. Note that in this particular case $H(z) = V(z)$ (see Subsection 2-4-1). The coefficients $a_k;\ k = 1, ..., p$ are called linear prediction coefficients (LPCs) and A is the filter gain, which is commonly assumed to be unity [21, 27, 69, 70]. Generally, the input is $\dot{u}[n]$, but we do not examine here the unvoiced case. In real speech signals, where the true order, p, of the vocal tract is unknown, Markel and Gray [71] proposed to set p slightly larger than the sampling frequency in kHz. For instance, if the sampling frequency is 8 kHz then a common choice is $p = 10$ [4]. Note also that SFM version 1 can be used in applications which aim to find a sparse residual. As we will see in Section 3-2, this can be achieved if we apply a pre-emphasis or a glottal-cancellation FIR filter prior to the analysis stage.

Assuming now that we want to estimate the poles of both the vocal tract and the glottal flow derivative, we use the SFM version 2

$$x[n] = \sum_{k=1}^{p+2} a_k x[n-k] + p[n] * r[n]. \tag{3-3}$$

The order-increment of 2 is based on the assumption that the glottal flow derivative has two poles outside the unit circle (see Section 2-2). The order-increment can also be set as

3 or 4, because the glottal flow derivative might also have one or two poles inside the unit circle, respectively, as was explained in Section 2-2. Now, the source signal is $p[n] * r[n] = p[n] - \alpha p[n-1]$. Thus, in each pitch period the source signal has two successive impulses; one positive which coincides on the corresponding impulse of the excitation, and one negative attenuated impulse at the next sample. The negative impulse is due to the zero of the lips which is not removed from the source in this SFM version. The SFM version 2 is more appropriate for sparse LP methods (see Subsections 3-3-5, 3-3-6).

Finally, when we want to estimate the excitation signal which is a quasi-periodic impulse train, we use the SFM version 3

$$x[n] = \sum_{k=1}^{p+2+M} a_k x[n-k] + p[n], \tag{3-4}$$

where M is the number of poles approximating the zero of the lips. The greater the M is, the better we approximate the zero. Here we select $M = 4$ as was proposed in [4]. Note that, the SFM versions 2 and 3 are mostly for applications that aim to produce a sparse residual. In order to avoid confusion in the next sections, we use the symbol p as the true order of the vocal tract. The order of LP is denoted by q and it can take the values p, $p+2$, or $p+6$, depending on the application.

LP methods estimate the LPCs through the minimization of the *prediction error sequence*, which is defined as

$$e[n] = x[n] - \widetilde{x}[n], \tag{3-5}$$

where $\widetilde{x}[n] = \sum_{k=1}^{p} a_k x[n-k]$ is the *linear predictor sequence* [4, 5]. Some LP methods try to minimize a weighted version of the prediction error sequence, i.e., $w[n] * e[n]$, where $w[n]$ is a weight function. The resulting minimum $e[n]$, say $\hat{e}[n]$, is called the residual. Note that, if the speech frame $x[n]$ is perfectly described by one of the SFM models of Equations 3-1, 3-3, 3-4, then the prediction error sequence is either $e[n] = \dot{u}_g[n]$, or $e[n] = p[n] * r[n]$, or $e[n] = p[n]$. However, SFM does not capture the non-linear interaction of the glottal flow with the vocal tract and also by using a small M in Equation 3-4 the zero is not estimated very well. Therefore, the true prediction error sequence is *approximately* given by the three aforementioned equations. For instance, if we select the SFM version 1 (i.e., Equation 3-1) as the input model for the LP analysis, we expect to see a ripple component on the residual (see Figure 2-5). This is because, LP cannot capture the time varying first formant of the vocal tract and all the non-linear interaction is transfered to the estimated glottal flow derivative signal (i.e., the residual). Moreover, in Figure 4-17, we see that the order increment of $M = 4$ does not cancel the zero of the lips completely in the residual. It just decrease the negative impulses which are one sample ahead of the positive impulses of the excitation signal (this can be observed better in the second row of Figure 4-17).

All LP analysis methods presented in the present thesis are special cases of a general LP problem which was proposed in [27, 72, 73]. The general LP problem is given by

$$[\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}] = \arg\min_{\boldsymbol{a}} \|\boldsymbol{W}\boldsymbol{e}\|_\gamma^\gamma = \sum_{n=N_a}^{N_b} |w[n]e[n]|^\gamma, \tag{3-6}$$

where $\boldsymbol{W}$ is a weighting diagonal matrix having as diagonal elements the sequence $\sqrt[\gamma]{w[n]}$, $\|.\|_\gamma$ is the L_γ norm, $\gamma > 0$, and $\boldsymbol{e}$ is the prediction error vector given by

$$\boldsymbol{e} = \boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a}. \tag{3-7}$$

Moreover, $\boldsymbol{x}_i$, $\boldsymbol{X}$ are given by

$$\boldsymbol{x}_i = [x[N_a - i] \ \ldots \ x[N_b - i]]^T,$$

$$\boldsymbol{X} = [\boldsymbol{x}_1 \boldsymbol{x}_2 \ \ldots \ \boldsymbol{x}_q],$$

respectively. If $N_a = 0$ and $N_b = N + q - 1$, the general LP problem 3-6 is set to the *autocorrelation mode*, while if $N_a = q$ and $N_b = N - 1$, is set to the *covariance mode* [18]. Furthermore, N_a and N_b may change in each speech frame if we set them to be the first and last points of the closed phase interval (see Subsection 3-3-2). The region $[N_a, N_b]$ will be called *analysis interval.* Note that the analysis interval is different from the *frame interval*, $[0, N - 1]$. We should mention here, that although the covariance mode may give more accurate estimates, it generally has an increased probability to produce unstable filters [4,18]. We should not forget to mention that, in the special case of the classical LP method (see Subsection 3-3-1) stability is guaranteed if it is set to the autocorrelation mode.

Usually, when the autocorrelation mode is selected for an LP method, the speech signal is windowed with a Hanning or a Hamming window before the analysis [4,65]. This is because the autocorrelation method uses some zero-samples outside the speech frame interval and, by using a Hanning window, the edges of the speech frame interval are smoothed in order to remove the sharp changes [4,65]. On the other hand, when the covariance mode is used, no windowing is performed (i.e., the window is rectangular), because in this mode no samples outside the speech frame interval are used [4,65]. In the present thesis, for ease of notation, we use in the sequel the same notation of the non-windowed speech frame for the windowed speech frame.

3-2 Pre-Emphasis & Cancellation of Glottal Flow Contribution

The classical LP analysis method [5,39] is a special case of the general LP problem 3-6 with configuration: $\boldsymbol{W} = \boldsymbol{I}$ and $\gamma = 2$. Therefore, it is equivalent to the linear least squares estimator (LLSE), which is given by

$$[\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}] = \arg\min_{\boldsymbol{a}} \|\boldsymbol{e}\|_2^2 = \arg\min_{\boldsymbol{a}} \|\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a}\|_2^2. \tag{3-8}$$

Using the Parseval Theorem [46], we can reformulate this minimization problem as

$$[\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}] = \arg\min_{\boldsymbol{a}} \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega, \tag{3-9}$$

where $E(e^{j\omega})$ is the discrete Fourier transform (DFT) of the prediction error sequence $e[n]$. Since $E(e^{j\omega}) = X(e^{j\omega})/H(e^{j\omega})$, Problem 3-9 is reformulated as

$$[\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}] = \arg\min_{\boldsymbol{a}} \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega = \arg\min_{\boldsymbol{a}} \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|X(e^{j\omega})|^2}{|H(e^{j\omega})|^2} d\omega. \tag{3-10}$$

This problem, referred to as the *spectral matching problem* [5], shows that the classical LP method tries to minimize the integrated ratio of the speech spectral magnitude to the spectral magnitude of the all-pole filter of order q. In other words, the classical LP method tries to fit $|H(e^{j\omega})|^2$ to $|X(e^{j\omega})|^2$. It can be proved that the ratio becomes unity (i.e., $|H(e^{j\omega})| = |X(e^{j\omega})|$), when $q \to \infty$ [5]. There are three possible disadvantages when LP analysis is applied on speech.

1. For small values of q (e.g., $q = p$), in problem 3-10, $|E(e^{j\omega})|^2$ is minimized more in the frequencies where the energy is higher. In inverse filtering applications, we have to set q equal to the true order, p, of the vocal tract (see Equation 3-1). In this case we have $E(e^{j\omega}) = \dot{U}_g(e^{j\omega})$. As was explained in Subsection 2-2-2, the spectral magnitude of the glottal flow derivative has high energy at the low frequencies (i.e., close to the glottal formant) and lower energy at the spectral tilt. Therefore, the minimization problem 3-10 fits better $|H(e^{j\omega})|^2$ to $|X(e^{j\omega})|^2$ in low frequencies, i.e., it estimates better the low-frequency formants than the high-frequency formants.
2. The estimated vocal tract will have the slope of $|X(e^{j\omega})|^2$ and not the slope of the true vocal tract, $|V(e^{j\omega})|^2$. The difference in the estimated slopes of $|X(e^{j\omega})|^2$ and $|V(e^{j\omega})|^2$ is clear in Figure 2-16.
3. The LP estimate, which is equivalent to the linear least squares estimate, is given by

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{x}_0. \tag{3-11}$$

   If we substitute $\boldsymbol{x}_0$ from Equation 3-7, Equation 3-11, becomes

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{a} + \boldsymbol{e}) = \boldsymbol{a} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{e}. \tag{3-12}$$

   In case $\boldsymbol{e}$ has zero mean, the estimate, $\hat{\boldsymbol{a}}$, will be unbiased, i.e., $E[\hat{\boldsymbol{a}}] = \boldsymbol{a}$. Otherwise, the estimate is biased. Assume now that the speech is perfectly characterized by Equation 3-1. Then, we may have biased estimates, because the glottal flow derivative $e[n] = \dot{u}_g[n]$ may not have a zero mean in some parts of speech (see Subsection 2-2-1). If Equation 3-3 perfectly describes the speech signal, we may still get a biased estimate, because $e[n] = p[n] - \alpha p[n]$, and α may be slightly less than 1. Finally, if Equation 3-4 describes perfectly the speech signal, it is much more likely to get a biased estimate, because $e[n] = p[n]$ does not have zero mean. Thus, if we obtain a biased estimate of LPCs, the estimated filter is given by

$$\hat{H}(z) = \frac{A}{1 - \sum_{k=1}^{p} \hat{a}_k z^{-k}} = \frac{A}{1 - \sum_{k=1}^{p} \left(a_k + \frac{(\boldsymbol{x}_k^T\boldsymbol{e})}{(\boldsymbol{x}_k^T\boldsymbol{x})}\right) z^{-k}}. \tag{3-13}$$

   Therefore, if the mean of $\boldsymbol{e}$ is not zero, then, on average, the estimated poles (i.e., the estimated formants) will not occur in the exact positions of the true poles of the vocal tract. We should emphasize here that the biased estimates are based completely on the assumption that the speech signal is perfectly described with one of Equations 3-1, 3-3, 3-4.

There are two general approaches of handling some or all of the aforementioned problems; the pre-emphasis methods, and the glottal flow cancellation methods. Both categories of methods try to remove the glottal pulse contribution from the speech signal, $x[n]$, before LP analysis is applied to it. Both categories of methods produce linear FIR filters that are applied to the speech signal prior to LP analysis (see Figure 1-1). The residual, $\hat{e'}[n]$, of the pre-emphasized or glottal-canceled signal, $x'[n]$, will not be anymore an estimate of the prediction error sequence $e[n]$, but it will be an estimate of the pre-emphasized or the glottal-canceled version of the prediction error sequence, $e'[n]$. This happens because of the SFM assumption in which the speech signal is generated by a cascade of LTI systems and, therefore, the FIR filter is combined with the prediction error sequence. An example of a pre-emphasized prediction error sequence, which is the glottal flow derivative, is shown in Figure 3-1.

3-2-1 Pre-emphasis

Usually, in order to boost the higher frequencies, a pre-emphasis filter is applied on the speech signal before LP analysis [4]. In inverse filtering applications, where we want to estimate the glottal flow derivative signal, the pre-emphasis filter has to be canceled after analysis by filtering the residual with the inverse of the pre-emphasis filter. On the other hand, in speaker localization, speech dereverberation, or speech coding applications[1], where we want a sparse residual, we do not have to cancel the pre-emphasis from the residual. This is because, pre-emphasis also increases sparsity as we can see in Figure 3-1. The most commonly used pre-emphasis filter [5, 28, 74–76] is given by

$$D_1(z) = 1 - \alpha z^{-1}, \tag{3-14}$$

where α typically ranges in the interval [0.96, 0.99]. This filter is identical to that used for modeling the lips, i.e., it is a high-pass filter which increases the slope of the spectral magnitude of speech by +6dB/octave. It is also an approximation of the first-order derivative. In words, if we apply the pre-emphasis filter of Equation 3-14 to the speech signal, before the minimization problem 3-10 is solved, then there is an increment of +6dB/octave to $|E(e^{j\omega})|^2 = |\dot{U}(e^{j\omega})|^2$. Although the +6dB/octave increase may solve the first [75] and the third problem, it is not enough to equalize the -12dB/octave roll-off of the spectral magnitude of the glottal flow derivative (when it has a non abrupt return phase) and, therefore, the second problem is not solved. Therefore, we need an additional +6dB/octave increase and, thus, a second-order pre-emphasis filter has to be used. This is given by

$$D_2(z) = \left(1 - \alpha z^{-1}\right)^2, \tag{3-15}$$

which approximates a second-order derivative. This pre-emphasis filter has been found to be very useful in estimating the glottal flow derivative signal when it is used in conjunction with the iteratively reweighted sparse linear prediction method presented in Subsection 3-3-6 [19, 77]. Therefore, the second-order pre-emphasis filter solves the first two aforementioned problems. We should mention at this point that when the glottal flow derivative signals have an abrupt closure, i.e., $t_e = t_c - 1$, a first-order pre-emphasis is enough to equalize the contribution of the spectral magnitude of the glottal flow derivative signal. This is because, when $t_e = t_c - 1$, the spectral magnitude of the glottal flow derivative signal has a -6dB/octave roll-off (see Subsection 2-2-2). So, to our opinion, the dominance of the first-order pre-emphasis filter in most inverse filtering applications might be caused by the fact that most papers published before the LF model based their assumptions on the Rosenberg model.

Now, lets see if the first-order or the second-order pre-emphasis filter handles efficiently the third problem. Since we have a cascade of linear systems, either the first-order or the second-order pre-emphasis filter can be convolved with the glottal flow derivative signal. If the glottal flow derivative does not have zero mean, its pre-emphasized version will have much smaller mean, especially so the second-order pre-emphasized version, as we can see in Figure 3-1. This means that both pre-emphasis filters help in the direction of obtaining a more unbiased estimate (i.e., more accurate formant locations).

Note, also in Figure 3-1, that the pre-emphasized version of the glottal flow derivative is sparser than the glottal flow derivative itself and, therefore, it makes the sparse linear

[1] In speech coding applications the removal of pre-emphasis may be performed at the reconstructed speech signal at the end.

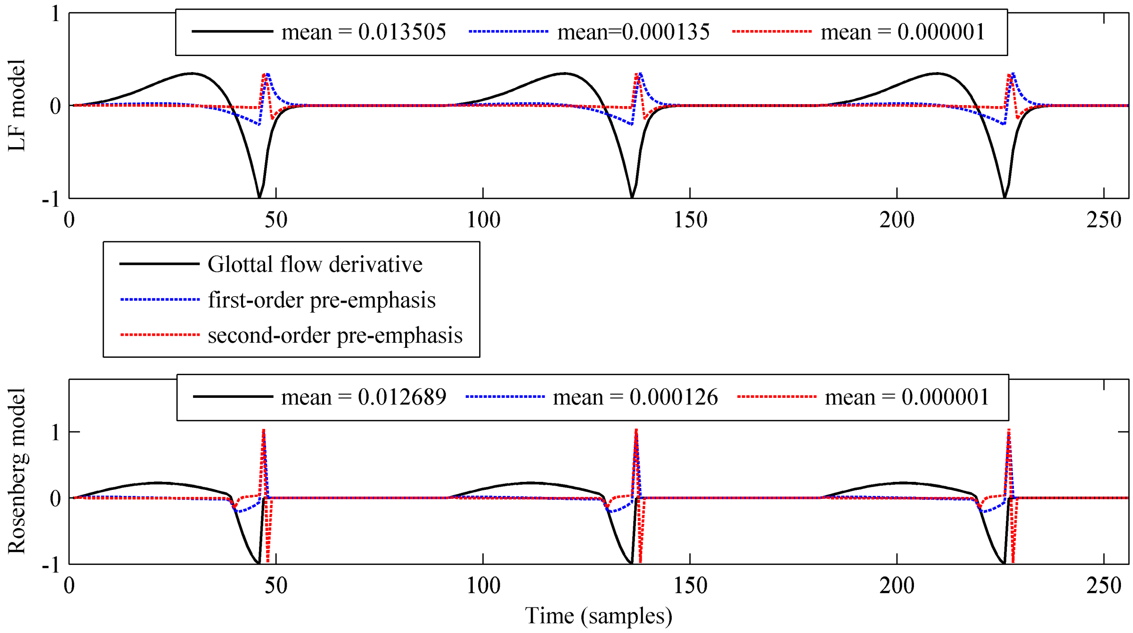


Figure 3-1: Mean values of the glottal flow derivative according to the LF and Rosenberg models, with first-order and second-order pre-emphasis.

prediction methods of Subsections 3-3-5 and 3-3-6 to be more appropriate than the classical LP method for the estimation of the LPCs and the glottal flow derivative [19]. Finally note, that the closed phase region increases in these simple pre-emphasized synthetic glottal flow derivative signals[2]. This make us to reasonably believe that pre-emphasis also helps the closed-phase analysis method to be less prone to errors caused by the non-accurate position of the covariance window. In addition pre-emphasis provides the latter method with the ability to perform better in cases of high pitch speakers where the closed phase interval is very short (see more about the closed-phase analysis method in Subsection 3-3-2).

3-2-2 Cancellation of Glottal Pulse

Although the pre-emphasis methods perform a kind of cancellation of the glottal pulse, they are distinguished in the present thesis from the methods that attempt to estimate and cancel the poles of the glottal pulse with zeros. In this subsection, we explore two different glottal pulse cancellation methods. The first one is part of the well known iterative adaptive inverse filtering (IAIF) method [22, 23], named IAIFGC, which is based on the classical LP method set to the autocorrelation mode. The second is a new method introduced in the present thesis and is based on the sparse LP method (see Subsection 3-3-5) set to the autocorrelation mode. We named this new technique sparse glottal pulse cancellation (SGPC) method. The main purpose of this new method is to show that pre-emphasis is, in general, more accurate than glottal-cancellation techniques based on IAIFGC framework for inverse filtering applications.

[2] In reality, a ripple component is also present in the opening part of the glottal flow derivative which cannot be removed by pre-emphasizing. Moreover, if the LF model has a discontinuity at GOI (i.e., at t_o), the pre-emphasized version of the modeled glottal flow derivative will have a peak at this instant.

IAIF method: The IAIF method is a complete glottal flow estimation method and has two versions. The first is pitch asynchronous [22], while the second is pitch synchronous [23]. In the present thesis we examine only the pitch asynchronous version, since SGPC is also pitch asynchronous. Specifically, the IAIF method consists of the following nine steps.

1. A rough first estimate of the glottal flow transfer function $U_g(z)$ is estimated through the classical LP analysis with order 1. The order is taken no greater than one in this step in order to avoid estimating some of the formants of the vocal tract.
2. The estimated $U_g(z)$ contribution is canceled from the speech signal, $x[n]$, via inverse filtering (i.e., filtering the speech signal with $1/U_g(z)$). The outcome of this procedure is denoted by $x_1[n]$.
3. A first estimate, $H_1(z)$, of the vocal tract transfer function is obtained through the classical LP analysis method of order p_1, applied to the signal $x_1[n]$.
4. The speech signal, $x[n]$, is filtered from $1/H_1(z)$ in order to obtain the first estimated glottal flow derivative signal, $\dot{u}_{g1}[n]$.
5. The estimated glottal flow derivative, $\dot{u}_{g1}[n]$, is integrated in order to obtain the corresponding glottal flow $u_{g1}[n]$. Usually, the integration is performed by the inverse of the filter of Equation 2-2.
6. The transfer function $U_{g1}(z)$ of the glottal flow, $u_{g1}[n]$, is estimated via the classical LP analysis method of order 4. To the author's opinion, the reason for selecting an order of 4 is that the method attempts to estimate the maximum-phase and minimum-phase parts of the glottal flow. These two parts are 3-4 poles in total (see Subsection 2-2-1).
7. The speech signal $x[n]$ is filtered with $1/U_{g1}(z)$ in order to remove the glottal flow contribution from it. The outcome of this process is denoted by $x_2[n]$.
8. The classical LP method of order p_2 is applied on $x_2[n]$ in order to estimate a better vocal tract transfer function, $H_2(z)$, than the estimated vocal tract transfer function, $H_1(z)$, of step 3.
9. The speech signal $x[n]$ is filtered with $1/H_2(z)$ in order to obtain a more accurate glottal flow derivative, $\dot{u}_{g2}[n]$, than the estimated glottal flow derivative of step 4, $\dot{u}_{g1}[n]$.
10. The estimated glottal flow derivative, $\dot{u}_{g2}[n]$, is integrated in order to obtain the final estimated glottal flow, $u_{g2}[n]$. Note that, in the present thesis we are interested in the estimation of the glottal flow derivative and, therefore, the final integration is not performed in our experiments and examples.

The orders p_1 and p_2 are selected as close as possible to the true order of the vocal tract. In [23] they were set equal to 10. Moreover, a linear-phase FIR high-pass filter with cut off frequency 30 Hz is applied on the speech signal, $x[n]$, before the whole procedure, in order to remove undesirable fluctuations of the estimated glottal flow [23]. In the present thesis, we will use the first seven steps of IAIF as a front end of all LP analysis methods of Section 3-3 and not only for the classical LP method as in IAIF. We call these seven steps the iterative adaptive inverse filtering glottal cancellation (IAIFGC) method in order to distinguish it from the whole IAIF method. To the author's knowledge IAIFGC is used only as a part of IAIF for inverse filtering applications. However in Chapter 4, we investigate its usefulness in sparse LP methods. This is because, after the glottal pulse cancellation from the prediction error sequence, we expect to obtain a residual which is closer to the excitation signal (i.e., a periodic impulse train).

SGPC method: In [19], the authors estimated very accurately the glottal flow derivative signal by combining the iteratively weighted sparse linear prediction method (see Subsection 3-3-6) with the second-order pre-emphasis filter $D_2(z)$. After trying many modifications of the IAIF framework, we observed that if we replace the first three steps of IAIFGC with this method, the performance of IAIF was increased in inverse filtering applications. We named this modification SGPC. In Chapter 4, we will see that, although with the SGPC modification we improve the performance of IAIF, we do not succeed higher estimation accuracy of the glottal flow derivative than the combination that SGPC is based, i.e., the sparse LP method combined with the second-order pre-emphasis filter. This interesting observation disputes the glottal-cancellation techniques based on the IAIFGC framework, since a simple pre-emphasis filter can achieve better results. Specifically, SGPC consists of the following 5 steps.

1. The speech signal $x[n]$ is pre-emphasized with the second-order pre-emphasis filter $D_2(z)$ (see Equation 3-15). The pre-emphasized speech signal is denoted by $x_1[n]$.
2. The sparse linear prediction method (see Subsection 3-3-5), set to the autocorrelation mode, with order $q = p$ (if the true order, p, of the vocal tract is known, otherwise we set $q = 10$) is applied on the pre-emphasized speech signal $x_1[n]$. The estimated filter is denoted by $H_1(z)$. Although the iteratively sparse linear prediction method can be used instead of the simple sparse linear prediction method, we observed that the estimation accuracy of the former is not significantly higher than the performance of the latter. The computational complexity can be reduced significantly by choosing the latter approach. Note that if there are any poles of the estimated filter, $H_1(z)$, lying outside the unit circle, they are replaced with their reciprocals.
3. The glottal flow derivative is estimated by filtering the non pre-emphasized speech signal $x[n]$ with the filter $1/H_1(z)$. The outcome is denoted by $\dot{u}_{g1}[n]$.
4. Similarly to the sixth step of the IAIF method, the transfer function of the glottal flow is estimated via the classical LP analysis method of order 4, but this time without any prior integration to $\dot{u}_{g1}[n]$. We observed that the integration deteriorates the performance of our method. The estimated transfer function is denoted by $U_{g1}(z)$. Note that this is the transfer function of the glottal flow and not of the glottal flow derivative because the transfer function of the glottal flow derivative possesses an extra zero that is not included in $U_{g1}(z)$.
5. The speech signal, $x[n]$, is filtered with $1/U_{g1}(z)$ and the outcome is denoted by $x2[n]$.

Unlike the IAIF method, the SGPC method does not use a linear-phase FIR high-pass filter at the beginning.

The estimation accuracy of the IAIF method is deteriorated significantly when the first formant of the vocal tract is very close to the glottal formant[3] [23], i.e., when it has a very low frequency. As we will see in the sequel, the second-order pre-emphasis combined with almost every LP method and SGPC combined with AM improve the estimation accuracy in such cases.

Similar to the single and double pre-emphasis, the FIR filters of IAIFGC and SGPC methods are convolved with the glottal flow derivative giving a sparser glottal-canceled version of the glottal flow derivative with a smaller mean than the glottal flow derivative itself (see Figure 3-2). Observe, also, that the closed phase region increases. Moreover, if a pre-emphasis

[3] As was discussed in Subsection 2-2-2, the glottal formant frequency usually takes values in the interval $[F_0/2, F_0/0.46]$.

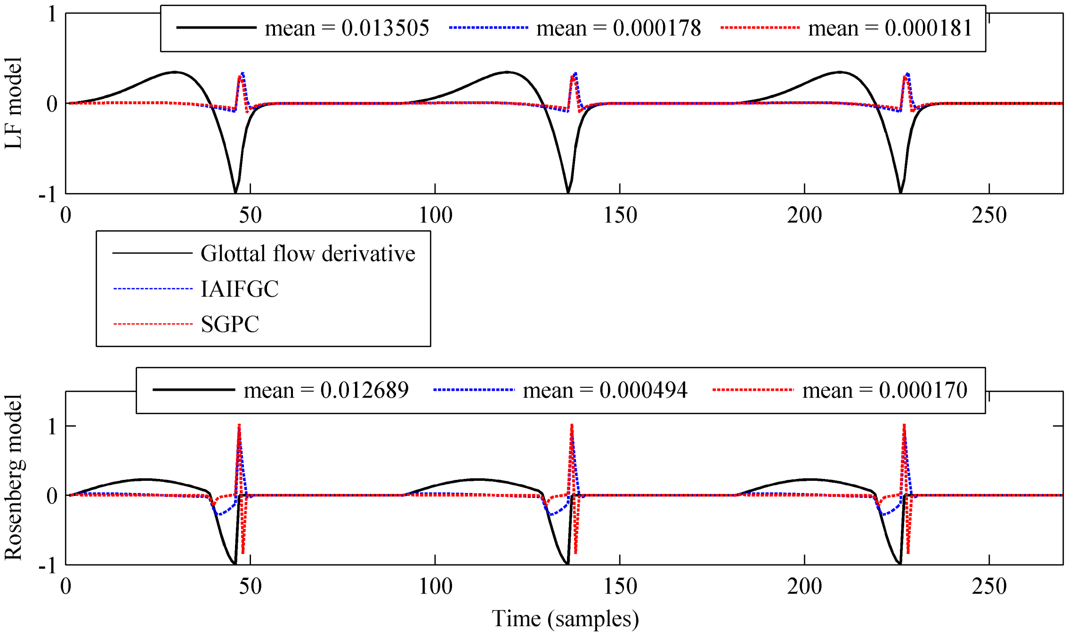


Figure 3-2: Mean values of the glottal flow derivative according to the LF and Rosenberg models, with IAIFGC and SGPC.

or a glottal-cancellation method is applied prior to LP, the SFM version 1 is applicable not only for inverse filtering but also for the generation of sparse residuals. This is because pre-emphasis and glottal-cancellation methods remove the glottal contribution from the prediction error sequence (which is the glottal flow derivative) leaving almost only the strong peaks of the periodic impulse train. Thus, in case of using a glottal-cancellation or a pre-emphasis technique prior to LP, an order increment may not improve the sparsity of the residual. On the contrary, it may deteriorate the degree of sparsity of the residual. For instance, if we filter a periodic impulse train with the filter $D_1(z)$, we obtain a new signal which is less sparse. All these issues will be examined in Chapter 4.

3-3 Linear Prediction Analysis Methods

In this section, seven LP analysis methods and their properties are reviewed. Note that some of these LP methods are popular for the estimation of the glottal flow derivative or the glottal flow signal, while others are popular for the estimation of a sparse residual. In Chapter 4, we will show that some methods that are well known for obtaining a sparse residual can be used also for finding a very accurate estimate of the glottal flow derivative, if they are properly combined with a pre-emphasis or a glottal-cancellation filter.

In inverse filtering applications, if we invert an unstable all-pole filter, the resulted inverse filter is not unstable anymore because it is an FIR filter which has all its poles at zero and, therefore, its ROC includes the unit circle. Nevertheless, it has been shown in [78] that the replacement of the estimated zeros, of the inverted estimated transfer function of the vocal tract, outside the unit circle with their reciprocals results in a more accurate estimation of the glottal flow derivative. As was discussed in Section 2-3, a simple and rational explanation for this improvement is that the vocal tract should be all-pole minimum-phase when the velum is closed (i.e., when the oral cavity works alone). More precisely, it is proven to be minimum-phase all-pole filter according to the Kelly-Lochbaum model. Therefore, for all LP methods, when an estimated pole of the vocal tract is outside the unit circle, it is replaced with its reciprocal.

3-3-1 Classical Linear Prediction Method

A brief introduction of the classical LP method has already been given in Section 3-2. Here we give a more details for this method for both the autocorrelation and covariance modes.

Autocorrelation method (AM): The autocorrelation mode of the classical LP method [5, 65, 71], which is referred in the literature as the autocorrelation method (AM), is one of the most well known and studied LP methods which gives very satisfactory results with very low complexity. AM is the a special case of the general LP problem with the configuration: $\boldsymbol{W} = \boldsymbol{I}$, $\gamma = 2$, $N_a = 0$ and $N_b = N + q - 1$, where q is the LP order and $\boldsymbol{I}$ the identity matrix. Therefore, AM can be written as the following LLSE problem:

$$\begin{aligned}[\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}] &= \arg\min_{\boldsymbol{a}} \|\boldsymbol{e}\|_2^2 = \arg\min_{\boldsymbol{a}} \|\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a}\|_2^2 \\ &= \arg\min_{\boldsymbol{a}} \| \begin{bmatrix} x[0] \\ \vdots \\ x[N+q-1] \end{bmatrix} - \begin{bmatrix} x[-1] & \cdots & x[-q] \\ \vdots & \ddots & \vdots \\ x[N+q-2] & \cdots & x[N-1] \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_q \end{bmatrix} \|_2^2.\end{aligned}$$

If we set the gradient of $\|\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a}\|_2^2$ with respect to $\boldsymbol{a}$ equal to zero, we obtain the normal equations

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{a}} = \boldsymbol{X}^T\boldsymbol{x}_0. \tag{3-16}$$

Thus, the linear least squares solution is

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{x}_0. \tag{3-17}$$

Note that in Equation 3-16 the components $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{x}_0$ are estimates of the autocorrelation matrix $\boldsymbol{R}_x$ and the autocorrelation vector $\boldsymbol{r}_x$. Therefore, we can reformulate Equation 3-16 as

$$\boldsymbol{R}_x\hat{\boldsymbol{a}} = \boldsymbol{r}_x, \tag{3-18}$$

or equivalently,

$$\begin{bmatrix} r_x(0) & r_x^*(1) & \cdots & r_x^*(q-1) \\ r_x(1) & r_x(0) & \cdots & r_x^*(q-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(q-1) & r_x(q-2) & \cdots & r_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{bmatrix} = - \begin{bmatrix} r_x(1) \\ r_x(2) \\ \vdots \\ r_x(q) \end{bmatrix}, \tag{3-19}$$

where the autocorrelation matrix $\boldsymbol{R}_x$ is Toeplitz. Thus, this linear system of equations can be solved very fast with the Levinson-Durbin algorithm [65].

Unlike the covariance method (see next paragraph), a very important property of AM is that it guaranties the minimum-phase property and, consequently, stability of the estimated filter [4]. Therefore, if the speech signal to be analyzed is mixed phase, the poles that lie outside the unit circle will be estimated as the reciprocals inside the unit circle [4]. Therefore, AM cannot estimate the true poles of the maximum phase component of the glottal flow. Note that the autocorrelation mode does not guarantee stability for every LP method but only for the classical LP method. However, compared to the covariance mode, it increases the probability of obtaining a minimum-phase estimated filter if it is applied to any LP analysis (see Section 3-1).

Covariance method (CM): One disadvantage of AM is that it forces the signal outside the interval $[0, N-1]$ to be zero. The covariance method (CM) [5, 65, 71] does not use samples outside this interval and, therefore, it does not need to force any samples to be zero. Thus, we do not have to use a Hanning window for the speech frame. Therefore, unlike AM, CM minimizes the prediction error sequence in a shorter time interval, $[N_a, N_b] = [q, N-1]$, to avoid using samples outside the interval $[0, N-1]$. Herein, CM is equivalent to the following LLSE problem:

$$\begin{aligned}[\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}] &= \arg\min_{\boldsymbol{a}} \|\boldsymbol{e}\|_2^2 = \arg\min_{\boldsymbol{a}} \|\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a}\|_2^2 \\ &= \arg\min_{\boldsymbol{a}} \| \begin{bmatrix} x[q] \\ \vdots \\ x[N-1] \end{bmatrix} - \begin{bmatrix} x[q-1] & \cdots & x[0] \\ \vdots & \ddots & \vdots \\ x[N-2] & \cdots & x[N-q-1] \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_q \end{bmatrix} \|_2^2.\end{aligned}$$

Similar to AM, if we take the gradient of $\|\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a}\|_2^2$ with respect $\boldsymbol{a}$ and set it to zero, we obtain the same normal equations as in Equation 3-16. However, this time the matrix $\boldsymbol{X}$ and the vector $\boldsymbol{x}_0$ are different from the corresponding quantities of AM. Thus, the autocorrelation matrix and the autocorrelation vector have now slightly different formulas. Specifically, the autocorrelation now is given by

$$r_x(k, t) = \sum_{n=q}^{N-1} x[n-t]x^*[n-k]. \tag{3-20}$$

Thus, using Equation 3-18, CM can be reformulated as

$$\begin{bmatrix} r_x(1,1) & r_x(1,2) & \cdots & r_x(1,q) \\ r_x(2,1) & r_x(2,2) & \cdots & r_x(2,q) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(q,1) & r_x(q,2) & \cdots & r_x(q,q) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{bmatrix} = - \begin{bmatrix} r_x(1,0) \\ r_x(2,0) \\ \vdots \\ r_x(q,0) \end{bmatrix}. \tag{3-21}$$

It can be observed that the autocorrelation matrix in not Toeplitz as in AM. Therefore, we can not use the Levinson-Durbin algorithm. The good news is that we can solve this problem with Cholesky decomposition and achieve lower complexity than Gaussian Elimination [65]. Finally note that, CM is usually preferred to AM when the frame sizes or the analysis intervals are very short [4].

3-3-2 Closed-Phase Covariance Method (CPCM)

The closed-phase covariance method (CPCM) [24, 78, 79] is used only in voiced speech and it takes advantage of the closed phase intervals of the glottal flow derivative. The main difference with CM is that it estimates the LPCs using as analysis interval the closed phase region only, i.e., $[N_a, N_b] = [t_c, t_o]$. In Section 2-2, we explained that in the closed phase region the source-filter interaction is very small or zero and, thus, the formants of the vocal tract are constant. Moreover, during the closed phase region the prediction error sequence is zero or approximately zero. Therefore, a more accurate estimate of the vocal tract transfer function can be obtained and, consequently, a more accurate estimate of the glottal flow derivative can be estimated via inverse filtering. Although this method estimates more accurately the glottal flow derivative and the vocal tract filter than the simple covariance method, it is more complex since we have to estimate also GCIs and maybe GOIs. We should not forget to mention that usually, a first-order pre-emphasis filter is applied on the speech signal prior to CPCM [24].

Note that one speech frame may contain multiple pitch periods and, therefore, multiple closed phases. However, CPCM usually uses as the analysis interval only one closed phase region. An exception is made in case of high pitch speakers (i.e., females and children). In these cases, if the closed phase is less than $q + 3$ samples, the Cholesky decomposition used for the analysis may fail [13]. In such cases, two closed phase intervals are used [13]. Note that, by using more than $2q$ samples of the closed phase interval for estimating the LPCs, the accuracy is not improved significantly [13]. This observation is very important because, as was explained in Chapter 2, the estimation of GOIs is a difficult task. Thus, we may estimate only the GCI positions and consider as the closed phase a number of samples after GCI, that will not be more than $2q$ or less than $q + 3$, hopping that the next GOI is after this number of samples.

We should not forget that pre-emphasis or glottal-cancellation can increase the closed-phase region if we assume that there is not the ripple component in the opening part of the glottal flow derivative. When a ripple component is present the situation is different, but still the pre-emphasis or glottal-cancellation may increase the robustness of CPCM with respect to the accuracy of *only*[4] the estimated GOI. The bad effects, caused by the inaccurate estimation of the GCIs and GOIs positions, can also be reduced by removing the poles located on the positive real axis [24, 78].

3-3-3 Weighted Linear Prediction Method (WLPM)

CPCM can give us very accurate estimates of the vocal tract filter and the glottal flow derivative signal, because it is applied on the closed phase intervals where the source-filter non-linear interaction is very small or zero. However, CPCM may face difficulties when the pitch frequency is relatively small, and its performance is strongly connected with the accuracy of GCI and GOI estimation algorithms.

The weighted linear prediction method (WLPM) [21], behaves somewhere between the classical LP method and CPCM, and tries to attain the good properties of both methods. First of all, it considers all the samples of the current speech frame as AM. Secondly, during the least squares minimization procedure, it applies a weight function on the speech frame. This weight function takes small values during the open phase and big values during the closed phase. Therefore, it gives more weight to the samples of speech that are not effected from the source-filter interaction, than to those that are effected more during the open phase.

[4] GCI positions do not change in the pre-emphasized or glottal-canceled versions of the glottal flow derivative (see Figures 3-1, 3-2).

Although this method is not so accurate as CPCM in case of long closed-phase intervals, it is less sensitive to pitch changes and is not dependent on the estimation of the GCI and GOI. However its performance is dependent on a parameter of the weight function as we will see in the sequel.

WLPM minimize the cost function

$$\varepsilon = \sum_{n=0}^{N+q-1} \left(x[n] - \sum_{k=1}^{q} a_k x[n-k] \right)^2 b[n]. \tag{3-22}$$

The weighting function $b[n]$ is given by

$$b[n] = \sum_{i=0}^{M-1} x[n-i-k]^2, \tag{3-23}$$

where k and M are scalars, usually taking values $k = 0, 1$ and $M \in [5, 15]$ if $f_s = 8$ kHz. In order to convert the minimization problem 3-22 to a special case of the general LP problem, we reformulate it as the weighted LLSE problem

$$[\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}] = \arg\min_{\boldsymbol{a}} \|\boldsymbol{B}^{1/2}\boldsymbol{e}\|_2^2 = \arg\min_{\boldsymbol{a}} \|\boldsymbol{B}^{1/2}(\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a})\|_2^2,$$

where $\boldsymbol{B}^{1/2}$ is a diagonal matrix having at its diagonal the weight function $\sqrt{b[n]}$. If we set the gradient of the cost function $\|\boldsymbol{B}^{1/2}(\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a})\|_2^2$ with respect to $\boldsymbol{a}$ equal to zero, we obtain

$$(\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X})\hat{\boldsymbol{a}} = \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{x}_0. \tag{3-24}$$

Therefore, the weighted LLSE solution is given by

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{x}_0, \tag{3-25}$$

where $\boldsymbol{B}$ is a diagonal matrix having at its diagonal elements the weight function $b[n]$ of Equation 3-23. Therefore, the weighted LP method is a special case of the general LP problem with configuration: $\boldsymbol{W} = \boldsymbol{B}^{1/2}$, $\gamma = 2$, and $N_a = 0$ and $N_b = N + q - 1$ (i.e., it is set to the autocorrelation mode).

Note that in Equation 3-24 the components $\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{x}_0$ are estimates of the weighted autocorrelation matrix $\boldsymbol{R}_x$ and the weighted autocorrelation vector $\boldsymbol{r}_x$, respectively. Therefore, we can re-write Equation 3-24 as

$$\boldsymbol{R}_x\hat{\boldsymbol{a}} = \boldsymbol{r}_x. \tag{3-26}$$

The weighted autocorrelation matrix is not Toeplitz anymore and, therefore, we cannot use the Levinson-Durbin method.

3-3-4 Iteratively Reweighted Least Squares Method (IRLSM)

The famous iteratively reweighted least squares method (IRLSM) [20] is a special case of the general LP method with configuration: $\gamma = 2$, $N_a = 0$, $N_b = N + q - 1$, $\boldsymbol{W} = \boldsymbol{B}^{1/2^{(k)}}$, where

$\boldsymbol{B}^{1/2^{(k)}}$ is a diagonal matrix with diagonal elements $\sqrt{b_{ii}^{(k)}}$ updated according to the Andrew's weight function [25] as

$$b_{ii}^{(k)} = \begin{cases} \frac{A^{(k)}\sin(\frac{\hat{e}^{(k-1)}[i]}{A^{(k)}})}{\hat{e}^{(k-1)}[i]}, & \text{if } |\hat{e}^{(k-1)}[i]| \leq \pi A^{(k)} \\ 0, & \text{if } |\hat{e}^{(k-1)}[i]| > \pi A^{(k)} \end{cases} . \tag{3-27}$$

In Equation 3-27, $\hat{\boldsymbol{e}}^{(k-1)}$ is the residual of the previous iteration and $A^{(k)}$ is the estimated scale computed as

$$A^{(k)} = \text{MAD}(\hat{\boldsymbol{e}}^{(k-1)})/0.6745, \tag{3-28}$$

where $\text{MAD}(\hat{\boldsymbol{e}}^{(k-1)})$ is the median absolute deviation of the residual of the previous iteration. Therefore, in iteration k, the optimization problem

$$[\hat{\boldsymbol{a}}^{(k)}, \hat{\boldsymbol{e}}^{(k)}] = \arg\min_{\boldsymbol{a}} \|\boldsymbol{B}^{1/2^{(k)}}\boldsymbol{e}\|_2^2 = \arg\min_{\boldsymbol{a}} \|\boldsymbol{B}^{1/2^{(k)}}(\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a})\|_2^2 \tag{3-29}$$

is solved. There are many weight functions that can be used in IRLSM. In the present thesis, we selected Andrew's weight function because we observed that it gives sparser residuals than many other weight functions presented in [20]. Note that, in the first iteration ($k = 0$), $\boldsymbol{B}^{1/2^{(0)}} = \boldsymbol{I}$. IRLSM terminates, when the algorithm converges. If the convergence needs a very high number of iterations, a lower maximum number of iterations is used as a stopping criterion.

3-3-5 Sparse Linear Prediction Method (SLPM)

The classical LP analysis method minimizes the variance (i.e., the squared L_2 norm) of the prediction error sequence and, therefore, is equivalent to the LLSE estimator. LLSE works well for unvoiced speech, because it is equivalent to the MLE estimator if the prediction error sequence is WGN [6]. On the other hand, in voiced speech, the prediction error sequence of Equations 3-3, 3-4 consists of quasi-periodic strong peaks. Also, the pre-emphasized or glottal-canceled version of the prediction error sequence of Equation 3-1 consists of quasi-periodic strong peaks (see Figures 3-1, 3-2). Therefore, LLSE suffers from outliers, i.e., it overemphasizes the large errors and puts less emphasis on smaller errors [5], producing a non-spiky residual (i.e., a non-sparse residual). Thus, in case of voiced speech, a desired property for an LP estimator is to estimate the LPCs such that the residual is sparse.

The sparse linear prediction method (SLPM) is a special case of the general LP problem with configuration $\boldsymbol{W} = \boldsymbol{I}$, $\gamma = 1$, $N_a = 0$, $N_b = N + q - 1$. Thus, we have the optimization problem

$$[\hat{\boldsymbol{a}}, \hat{\boldsymbol{e}}] = \arg\min_{\boldsymbol{a}} \|\boldsymbol{e}\|_1 = \arg\min_{\boldsymbol{a}} \|\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a}\|_1. \tag{3-30}$$

This optimization problem is also known as the LAD estimator [80]. To the author's knowledge, the application of the LAD estimator in speech was first explored in [69]. In particular, the authors of [69] used the Burg algorithm in order to solve the L_1-problem. Although the minimum-phase property of this algorithm has been proven, it behaves somewhere in between the L_2 and the L_1 minimization in terms of sparsity [18]. For this reason, in [18, 27] the L_1-problem is solved with an interior-point algorithm which gives sparser residuals but it does not guarantee the minimum-phase property of the estimated all-pole filter [27].

3-3-6 Iteratively Reweighted Sparse Linear Prediction Method (IRSLPM)

In [81], it was shown that the re-weighting of the L_1 minimization enhances the sparsity of the residual. Thus, the authors of [77] applied this theory to speech and obtained sparser residuals compared to the corresponding ones of the sparse linear prediction method. The configuration of the general LP problem for the iteratively reweighted sparse linear prediction method (IRSLPM) is $\gamma = 2$, $N_a = 0$, $N_b = N + q - 1$, $\boldsymbol{W} = \boldsymbol{B}^{(k)}$, where $\boldsymbol{B}^{(k)}$ is a diagonal matrix with diagonal elements $b_{ii}^{(k)}$ updated as

$$b_{ii}^{(k)} = \frac{1}{|\hat{e}^{(k-1)}[i]| + c}. \tag{3-31}$$

In Equation 3-31, $\hat{\mathbf{e}}^{(k-1)}$ is the residual of the previous iteration and the constant c is selected to be in the order of the expected nonzero magnitude of $\boldsymbol{e}$ [81]. Therefore, IRSLPM is given by the iterative optimization problem

$$[\hat{\boldsymbol{a}}^{(k)}, \hat{\boldsymbol{e}}^{(k)}] = \arg\min_{\boldsymbol{a}} \|\boldsymbol{B}^{(k)}\boldsymbol{e}\|_1 = \arg\min_{\boldsymbol{a}} \|\boldsymbol{B}^{(k)}(\boldsymbol{x}_0 - \boldsymbol{X}\boldsymbol{a})\|_1. \tag{3-32}$$

Note that, the authors of [77] used a constant c for their experiments. In the present thesis, we adaptively change the parameter c as

$$c = \max\left(\frac{2}{N}\sum_{n=0}^{\lceil (N-1)/2 \rceil} \hat{e}_{\mathrm{d}}^{(k-1)}[n], 0.001\right), \tag{3-33}$$

where $e_{\mathrm{d}}[n]$ is the sorted version of the residual of the previous iteration in decreasing order. This adaptive formula of c is very similar to the corresponding adaptive formula that was used in [81]. We observed that by adaptively changing c according to Equation 3-33, the sparsity is increased slightly on average, over multiple frames, compared to a constant selection of c.

Furthermore, note that, in the first iteration ($k = 0$) $\boldsymbol{B}^{(0)} = \boldsymbol{I}$. IRSLPM is solved iteratively until $\|\boldsymbol{e}\|_1$ becomes smaller than a threshold. If this requires a high number of iterations, a lower maximum number of iterations is used as the stopping criterion. Finally, we should remind that this is the procedure used by the SGPC method for estimating and canceling the poles of the glottal flow from the speech signal.

3-3-7 Weighted Epoch Linear Prediction Method (WELPM)

The weighted epoch linear prediction method (WELPM) [70] has exactly the same configuration as WLPM with only difference in the weight function. This is now given by

$$b[n] = 1 - \sum_{k=1}^{N_{t_e}} g[n - n_k], \tag{3-34}$$

where n_k are the t_e positions in $x[n]$ which are determined via the SEDREAMS algorithm [24]. The function $g[n]$ is a Gaussian function, i.e., $g[n] = \kappa e^{(-n/\sigma)^2}$. The constant k is set slightly less than 1, e.g., $\kappa = 0.9$ [70] and the standard deviation depends on the pitch period and sampling frequency. This weight function de-emphasizes the large residual errors in t_e instants resulting in a more robust and sparse estimate than the classical LP method. The performance of this method is strongly connected to the SEDREAMS estimation accuracy of the t_e instants, and to the proper choice of the parameters σ and κ.

Chapter 4

Experiments

In this chapter we evaluate the LP methods combined with all pre-emphasis and glottal-cancellation methods discussed in Chapter 3. Our aim is to find the best combination for inverse filtering and speech coding applications. The combinations are denoted by triplets (method, filter, order), where the field method is one of the LP methods, the field filter is one of the four pre-emphasis/glottal-cancellation methods or nothing and the order can be p, $p+2$ or $p+6$. The rest of this chapter is organized as follows. The evaluation methodology and measures are introduced in Section 4-1. In Section 4-2, we provide an experimental evaluation of several combinations in terms of spectral magnitude estimation accuracy, glottal flow derivative estimation accuracy, sparsity of the residual, percentage of stable estimated filters and robustness to reverberation. We also give some examples of synthetic and true speech signals. Finally, in Section 4-3 we compare the performances of the best performing combinations of Section 4-2 in the context of speech dereverberation.

4-1 Evaluation Methodology and Measures

To evaluate a speech analysis method, we use the Gini index measuring the sparsity of the residual, the log spectral distortion distance (LSD) metric measuring the spectral distortion between true and estimated vocal tracts, and the signal to noise ratio (SNR) measuring the accuracy of the estimated glottal flow derivative signal with respect to the true glottal flow derivative signal.

Sparsity: Any discrete-time signal, $y[n]$; $n = 0, ..., N-1$, can be decomposed as the product of an $N \times M$ ($M \geq N$) dictionary $\boldsymbol{\Psi}$ and an $M \times 1$ vector $\boldsymbol{r}$, i.e.,

$$\boldsymbol{y} = \boldsymbol{\Psi}\boldsymbol{r}, \tag{4-1}$$

where $\boldsymbol{\Psi}$ can be, for example, the $N \times N$ inverse Fourier transform matrix and r the frequency vector. If we select a dictionary $\boldsymbol{\Psi}$ such that the signal $\boldsymbol{y}$ can be well-approximated as linear combinations of just a few column vectors (atoms) of $\boldsymbol{\Psi}$, we say that $\boldsymbol{r}$ is sparse [81]. This dictionary may be the $N \times N$ identity matrix and, thus, the signal is sparse in the time domain.

In the present thesis, we are interested in the sparsity of the residual, $\hat{\boldsymbol{e}}$. Sparsity can be measured with various metrics [16]. The L_0 quasi-norm is one such metric and is defined as $\|\hat{\boldsymbol{e}}\|_0 = |\{i, \hat{e}_i \neq 0\}|$, which is the number of non-zero elements of the vector $\hat{\boldsymbol{e}} = [\hat{e}[0], \hat{e}[1], ..., \hat{e}[M-1]]^T$. The Gini index [16] is another sparsity measure that satisfies all the desirable properties for a sparsity measure [16] (i.e., *Robin Hood*, *scaling*, *rising tide*, *cloning*, *Bill Gates* and *babies*), and is defined as

$$\text{Gini}(\hat{\boldsymbol{e}}) = 1 - 2\sum_{m=1}^{M} \frac{\hat{e}_o(m)}{\|\hat{\boldsymbol{e}}_o\|_1}\left(\frac{M - m + 1/2}{M}\right),$$

where $\hat{\boldsymbol{e}}_o$ is the sorted version of $\hat{\boldsymbol{e}}$ in increasing order. The Gini index takes values in the interval [0,1). The previous two sparsity measures, however, are not convex and are not efficiently optimized in a sparse all-pole analysis optimization problem. Instead, the L_1-norm is often used [27, 69, 82], which is an approximation of the L_0 norm. The estimator that minimizes the L_1-norm of the prediction error sequence is called the least absolute deviations (LAD) estimator and is equivalent to the maximum likelihood estimator (MLE) estimator when the prediction error sequence follows a Laplacian distribution [80].

Now, lets see in more detail the seven desirable properties mentioned before and identify those that are satisfied by each of the aforementioned sparsity measures [16]. The Robbin Hood criterion states that, if the large elements of $\hat{\boldsymbol{e}}$ give some of their energy to the smaller elements, the sparsity is decreased. The scaling criterion states that, if we multiply with a constant all the elements of $\hat{\boldsymbol{e}}$, the sparsity remains the same. The rising tide criterion states that, if we add a constant to each element of $\hat{\boldsymbol{e}}$, the sparsity is decreased. The cloning criterion states that, if we concatenate two vectors with the same elements (i.e., $[\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}^T]^T$), the sparsity remains the same. The Bill Gates criterion states that, if one individual element becomes infinitely large, sparsity is increased. Finally, the babies criterion states that, if we add extra zero elements to $\hat{\boldsymbol{e}}$, sparsity will increase. The L_0 norm satisfies only the scaling and babies criterion, the L_1 norm satisfies only the rising tide criterion, while the Gini index satisfies all the above six criteria. Although we used the L_1 norm in some of the convex optimization problems of Chapter 3, the Gini index is used for the evaluation in this chapter for three reasons. First of all, in this chapter some of the LP methods that we evaluate are not based on L_1 minimization but on L_2 minimization. Secondly, the Gini index will give us a more objective view of sparsity for both L_1 and L_2 methods. Thirdly, our results will be independent of the frame size, the sampling frequency and the amplitude of the residual, because of the aforementioned properties of the Gini index. Thus, they will be comparable with previously and future published papers that use the Gini index.

Spectral distortion measures: In the present thesis, we are interested to measure the spectral distortion between the estimated power spectrum of the vocal tract, $P_{\hat{\boldsymbol{h}}}(w_m) = |\hat{H}(e^{j\omega})|^2$, and the true power spectrum of the vocal tract, $P_{\boldsymbol{h}}(w_m) = |H(e^{j\omega})|^2$, where $h[n]$ is the impulse response of the vocal tract system.

Itakura-Saito distance (ISD) [69, 83] is a distortion measure given by

$$\text{ISD}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{P_{\boldsymbol{h}}(w_m)}{P_{\hat{\boldsymbol{h}}}(w_m)} - \log\frac{P_{\boldsymbol{h}}(w_m)}{P_{\hat{\boldsymbol{h}}}(w_m)} - 1\right), \tag{4-2}$$

where M is the fast Fourier transform (FFT) size. ISD is non-symmetric (i.e., $\mathrm{ISD}(\boldsymbol{h}, \hat{\boldsymbol{h}}) \neq \mathrm{ISD}(\hat{\boldsymbol{h}}, \boldsymbol{h})$) and, therefore, it is not a metric. ISD is used in the Linde-Buzo-Gray (LBG) algorithm [84] as the distortion measure for the construction of the linear prediction coefficients (LPC)s codebook [4]. Compared to other distortion measures, the usage of ISD in LBG is a good choice in terms of the perceptual quality of the synthesized speech [84]. As was explained in Section 2-3, the spectral peaks are perceptually more important than the spectral valleys, and this distinction is included in the ISD measure.

Another famous distortion measure is the COSH distance [17], which is a symmetric version of ISD and satisfies all the other properties of a proper metric. COSH is given by

$$\mathrm{COSH}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \frac{1}{2}\left(\mathrm{ISD}(\boldsymbol{h}, \hat{\boldsymbol{h}}) + \mathrm{ISD}(\hat{\boldsymbol{h}}, \boldsymbol{h})\right). \tag{4-3}$$

Finally, in the present thesis we use the simple distortion measure

$$\mathrm{LSD}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(10\log_{10}\frac{P_{\boldsymbol{h}}(w_m)}{P_{\hat{\boldsymbol{h}}}(w_m)}\right)^2}, \tag{4-4}$$

which is is the log spectral distortion distance (LSD) and is a proper metric [17].

Evaluation of the estimated glottal flow derivative: The SNR measure is used to evaluate the accuracy of the estimated glottal flow derivative, $\hat{\dot{\boldsymbol{u}}}_g$, relative to the true one, $\dot{\boldsymbol{u}}_g$. It is given by

$$\mathrm{SNR} = 10\log_{10}\frac{\|\dot{\boldsymbol{u}}_g\|_2^2}{\|\dot{\boldsymbol{u}}_g - \hat{\dot{\boldsymbol{u}}}_g\|_2^2}. \tag{4-5}$$

In inverse filtering applications we are interesting in the structure differences of the estimated and the true glottal flow derivatives. Sometimes, when the order of the estimated filter is not the same with the true order of the vocal tract, we may have differences in the amplitude gains but not in the structures of the estimated glottal flow derivative and the true one. In order to be able to compare with SNR the structures of the two waveforms, a least squares normalization/projection is performed on the estimated glottal flow derivative with respect to the true one before the computation of SNR.

Synthetic Speech for Evaluation: In real speech signals we do not know the exact form of the true glottal flow/glottal flow derivative and the true vocal tract and, therefore, we cannot evaluate the performance of the LP analysis methods in terms of the estimation accuracy. Therefore, synthetic signals have to be used. A rarely used methodology [78] to tackle this problem is to generate synthetic glottal flow signals with the lumped-element model considering also the interactions of the glottal flow with the supraglottal and subglottal parts. This whole procedure is implemented in the LeTalker software [61]. Moreover, LeTalker is capable of estimating the vocal tract transfer functions using the Story model (see Section 2-3).

It was shown in Subsection 2-2-1 that the lumped-element model is problematic in modeling smooth openings and closings of the glottal flow (i.e., it cannot model the return phase). Therefore, in the present thesis, we use a source-filter model to generate the synthetic speech signals as in [23,85]. In particular, we use the LF model for the source. This gives us the flexibility to test various glottal flow derivative signals with different parameterizations. We also use the vocal tract impulse responses, generated by LeTalker according to the Story model. These are generated using the area functions of a male speaker, acquired via magnetic resonance imaging (MRI) [34]. From the impulse responses we construct the vocal tract transfer functions $H(z)$ with gain $A = 1$. Note that the synthetic signals do not contain the non-linear interactions of the glottal flow with the supraglottal and subglottal parts as in [78].

4-2 Evaluation

In this section, we provide two different experimental evaluations of the LP methods. In Subsection 4-2-1, we provide an evaluation of several combinations in terms of the estimation accuracy of the vocal tract spectral magnitude and the glottal flow derivative, using synthetic signals. For this evaluation we use several synthetic signals parametrized with various ways and we compute the average values and standard deviations of SNRs and LSDs. We also show how much the estimation accuracy is degraded under reverberation phenomena. Moreover, we give some inverse filtering examples of synthetic and true speech signals without reverberation. In Subsection 4-2-2, we provide an experimental evaluation of the LP methods in terms of sparsity and stability using real speech signals with and without reverberation. Unlike in inverse filtering evaluation, in Subsection 4-2-2 we can use real speech signals, because the assessment of sparsity and stability does not need the knowledge of the true glottal flow derivative and the true vocal tract.

4-2-1 Inverse Filtering

The estimation accuracy of the glottal flow derivative and the vocal tract is assessed using the SNR and LSD measures, respectively. For this purpose, we test all possible combinations, with the same LP order set to the true order of the vocal tract, i.e., $p = 8$. Therefore, for ease of notation, we omit the last field of the triplets used to denote the combinations. We used 270 different LF glottal flow derivative signals and 8 different vocal tract transfer functions producing the vowels /i/, /ae/, /uh/, /ah/, /aw/, /oh/, /U/ and /u/. Therefore, we used $270 * 8 = 2160$ synthetic signals in total. In particular, the glottal flow derivatives are parametrized as follows. All the elements of the set $\{5, 6.25, 7.5, 8.75, 10, 11.25\}$ are tested as the pitch period. All the elements of the set $\{0.4, 0.5, 0.6, 0.7, 0.8\}$ are tested as the open quotient. All the elements of the set $\{0.7, 0.8, 0.9\}$ are tested as the asymmetry coefficient. All the elements of the set $\{0.05, 0.10, 0.15\}$ are tested as the R_a parameter. In all LF parametrizations the parameter E_e was the same and it was set to -1. In the LF model we assumed for convenience that $t_c = T_0$. Since we do not have an ideal closed phase region, we assume that the closed phase region is the interval $[t_e + t_a, t_e + t_a + p + 3]$ (for explanation see Subsections 2-2-1, 3-3-2). Since the closed phase region is not ideal, we expect to see a moderate performance of CPCM. Finally, the speech frame is 40 ms and the sampling frequency is 8 kHz.

No reverberation: Figures 4-4 - 4-13 show the SNR and LSD average values and the standard deviations for each combination. Each figure has two sub-figures. The top sub-figure shows

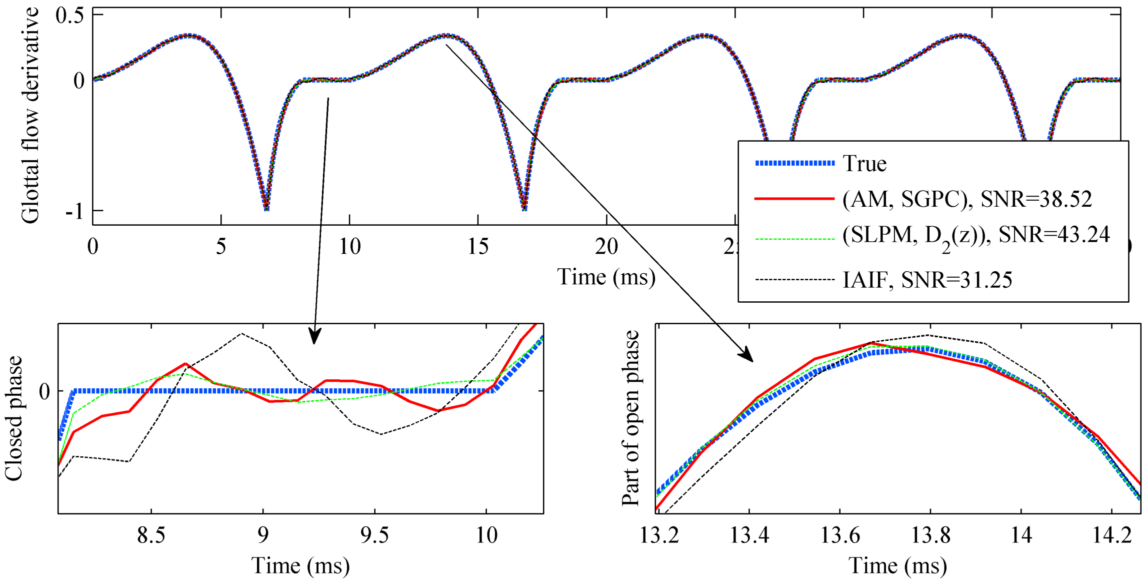


Figure 4-1: Inverse filtering example for the vowel /ah/ with various combinations.

the SNR average values and standard deviations, while the bottom sub-figure shows the LSD average values and standard deviations. Note that in the cases of CPCM, WLPM, WELPM there are two figures. In the first figure of CPCM, the analysis interval is only $p + 3$, while in the second figure the analysis interval is longer (i.e., $p + 20$). Both analysis intervals are between two successive t_e instants for every pitch period.

Figure 4-14 consists of the best performing combinations of each LP method. As we can see, the best combination in terms of inverse filtering accuracy is (SLPM, $D_2(z)$). Note that this combination is similar to the method that is proposed in [19] for inverse filtering. The only difference is that in [19] the proposed method is IRSLPM. Although IRSLPM finds sparser residuals (see Subsection 4-2-2) than SLPM, it is slightly worse than SLPM for inverse filtering purposes as seen in Figure 4-14. Moreover, we should notice that (AM, SGPC) improves performance compared to the IAIF method, but it is worse than (SLPM, $D_2(z)$). Therefore, by using (SLPM, $D_2(z)$) as part of SGPC (see Subsection 3-2-2) the performance is not improved compared to (SLPM, $D_2(z)$). On the contrary, the resulting method is less accurate and more complex. This means that the third and fourth steps of SGPC, or the fourth and sixth steps of IAIFGC do not increase performance. This is an interesting observation, because it disputes the accuracy and the methodology of the general glottal-cancellation framework of IAIFGC. Therefore, it would be interesting to investigate, in the future, if the combination (SLPM, $D_2(z)$) is more accurate than other glottal-cancellation methods which are based on the IAIFGC framework (e.g., [86]). Finally, as was explained in Section 3-2, CPCM becomes more robust if it is combined with a pre-emphasis or glottal-cancellation technique, in the case of a longer analysis interval (in the direction of GOI) than the true closed phase region.

Figures 4-1 and 4-2 show two inverse filtering examples, with some combinations, for the vowels /ah/ and /i/, respectively. Both vocal tract transfer functions have 4 formants and, therefore, $p = 8$. The frequencies of the first formant of the /ah/ and the /i/ vowels are 802 Hz and 228 Hz, respectively. The gain of the vocal tract filters are set to unity. The glottal

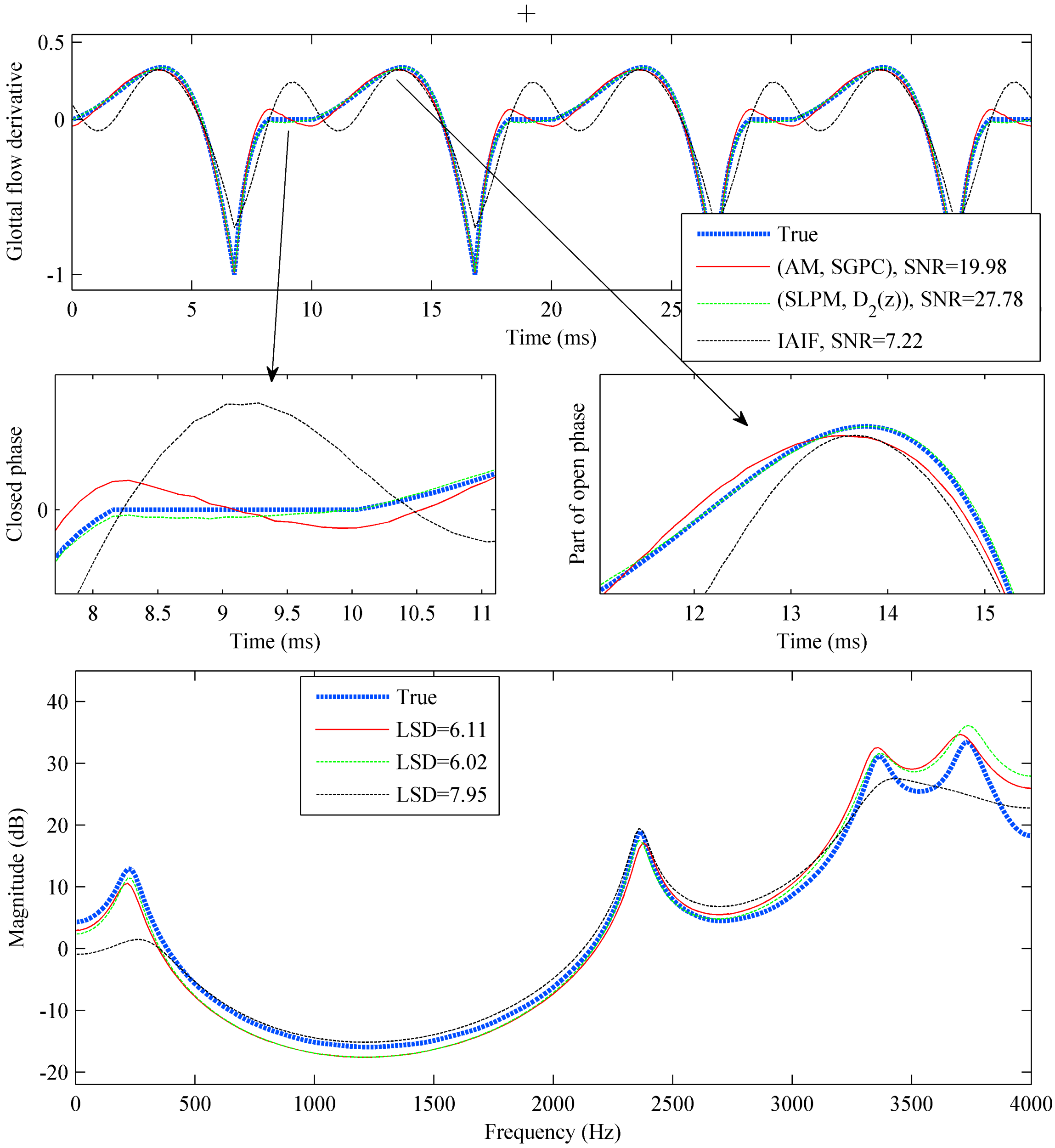
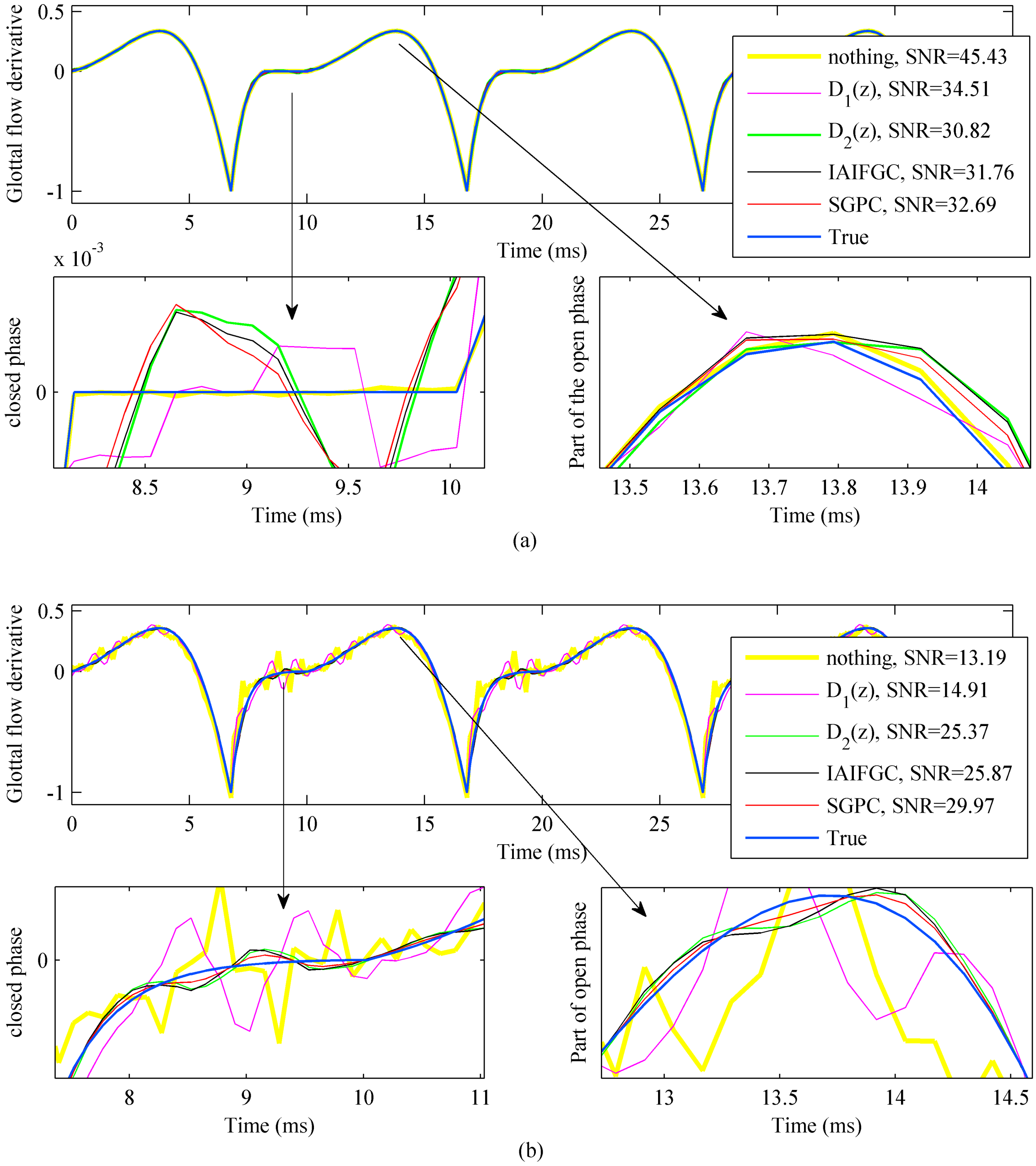


Figure 4-2: Inverse filtering example for the vowel /i/ with various combinations.

flow derivative signal has pitch period 10 ms and its parametrization is: $E_e = -1$, $O_q = 0.675$, $a_m = 0.7962$, $R_a = 0.05$, $R_c = 0.0875$. The glottal formant frequency is $f_p = 74.07$ Hz and, therefore, it is very close to the first formant of the vocal tract of the /i/ vowel. When the first formant of the vocal tract is very close to the glottal formant, the inverse filtering procedure becomes a difficult task [23], but as seen in Figure 4-2, the combination (SLPM, $D_2(z)$) performs very well. We should mention that we have omitted to include in Figure 4-1 the spectral magnitudes of the estimated and the true vocal tract, because they are almost identical and the difference is difficult to be observed.

If we have an ideal closed-phase region (i.e., when $t_c < T_0$ and $T_0 - t_c$ is greater than

$p+3$), the performance of CPCM improves significantly. Figure 4-3 shows the performance of CPCM when the vocal tract produces the vowel /ah/ for two glottal flow derivatives. The first glottal flow derivative (see the bottom sub-figure) is one of the glottal flow derivatives used in our experiments. This glottal flow derivative does not have an ideal closed phase region (remember that the LF model was constructed with $t_c = T_0$) and the analysis interval is $[t_e + t_a,\ t_e + t_a + 15]$. The second glottal flow derivative (see top sub-figure) has an ideal



(a)

(b)

Figure 4-3: Inverse filtering example for the vowel /ah/ with various combinations of CPCM with an ideal closed phase region (a) and a non-ideal closed phase region (b).

closed phase region with length $T_0 - t_c = 15$ samples, which is greater than $p + 3$ (see more details about the number $p+3$ in Subsection 3-3-2). In the latter case, the analysis interval is the same with the closed phase region. As seen in Figure 4-3, the latter case performs much better. Note also that in the first case, (CPCM, SGPC), (CPCM, IAIFGC) and (CPCM, $D_2(z)$) improves performance and, consequently, robustness of the CPCM method. To the author's experience, most of the times the glottal flow derivative does not have an ideal closed

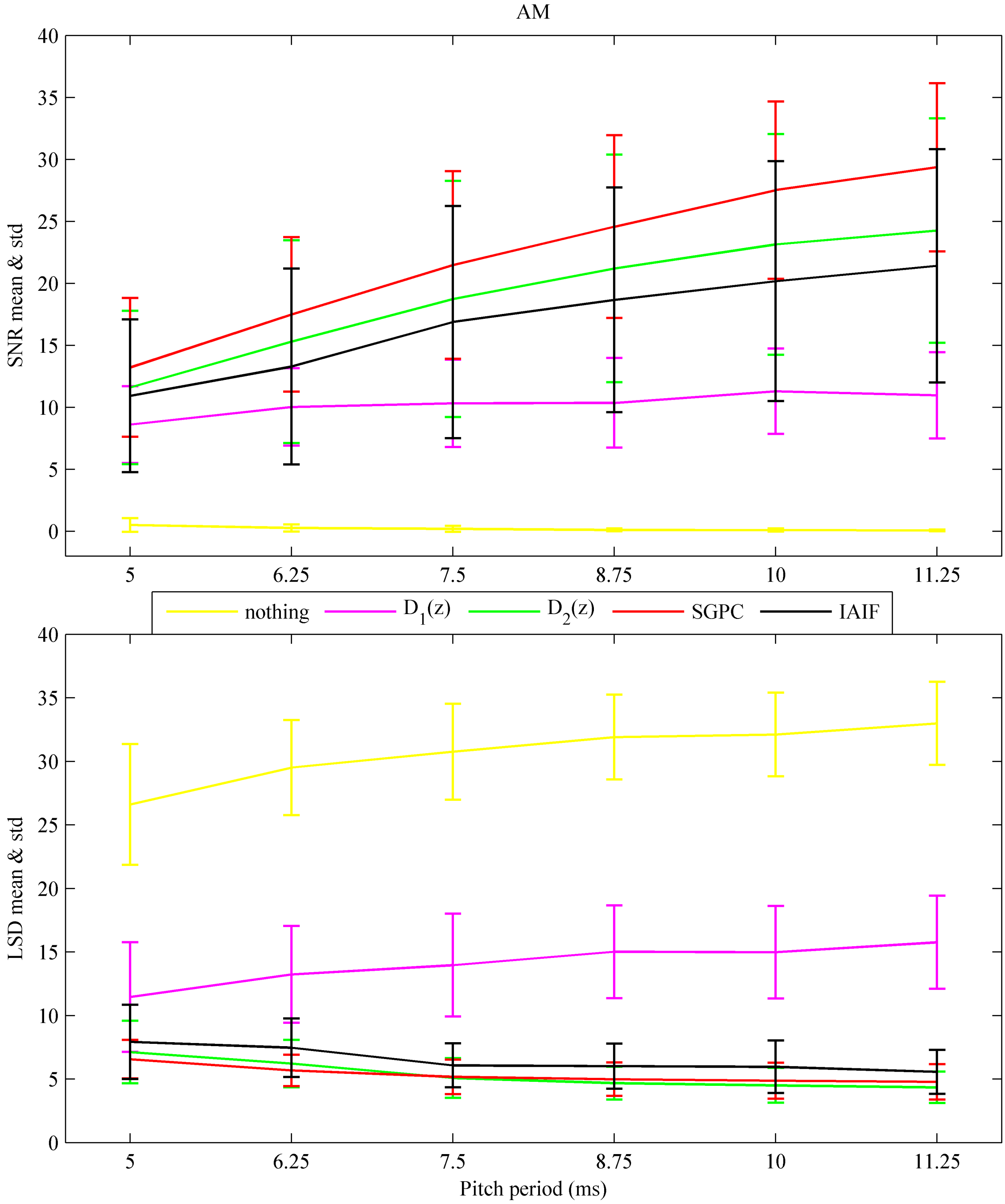


Figure 4-4: AM: means and standard deviations of SNRs and LSDs.

phase region.

Finally, Figure 4-15 shows the estimated glottal flow derivative and glottal flow waveform using various combinations for a real speech signal. As was expected, the combination (AM, $D_1(z)$) does not perform well. Although we cannot measure the SNR of this combination,

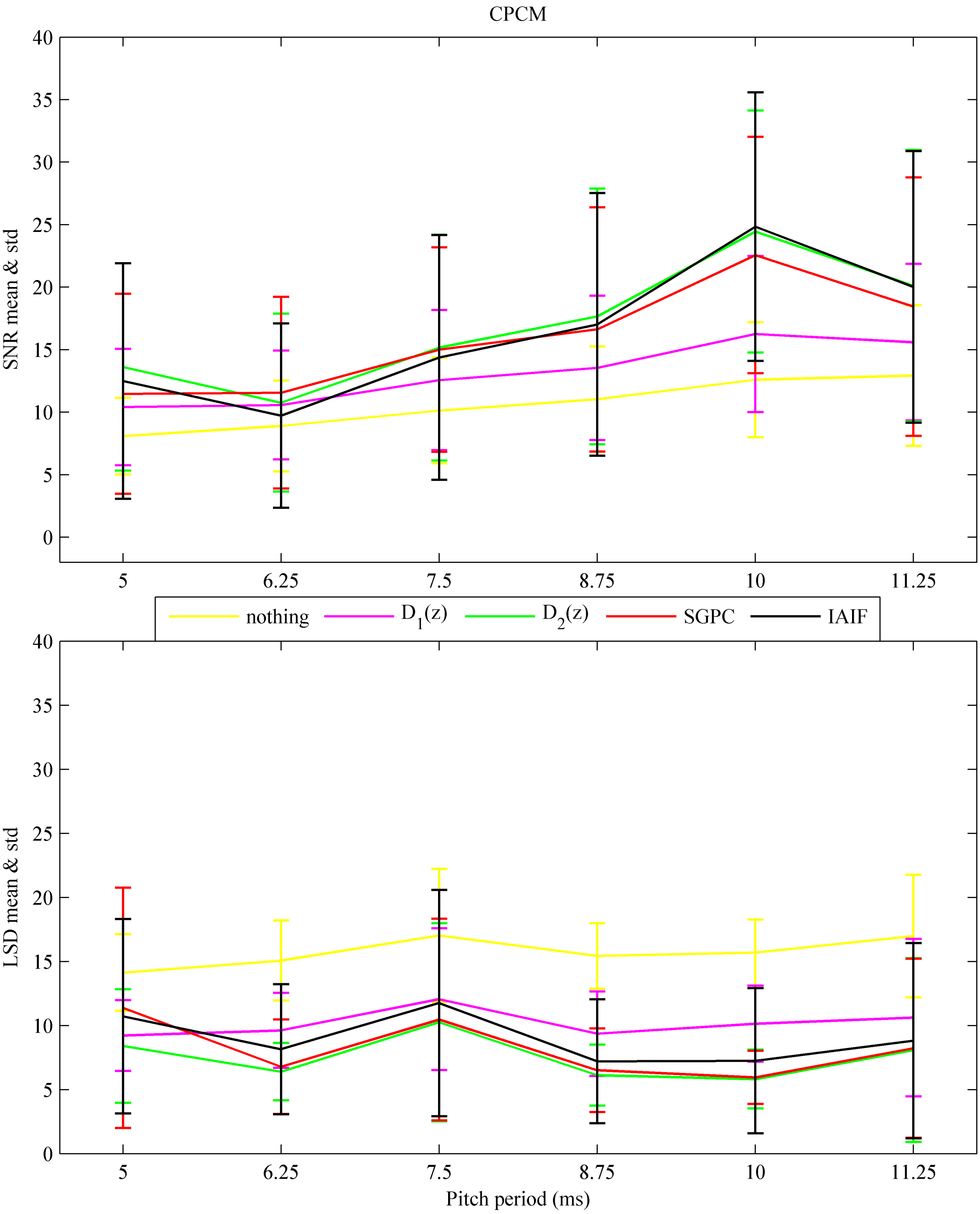


Figure 4-5: CPCM, (analysis interval= $p + 3$): means and standard deviations of SNRs and LSDs.

we can observe the structure of the estimated glottal flow derivative and see that the closed phase region is not estimated well and it is much more noisy than all the other estimated glottal flow derivatives. The differences of all the other methods are very small.

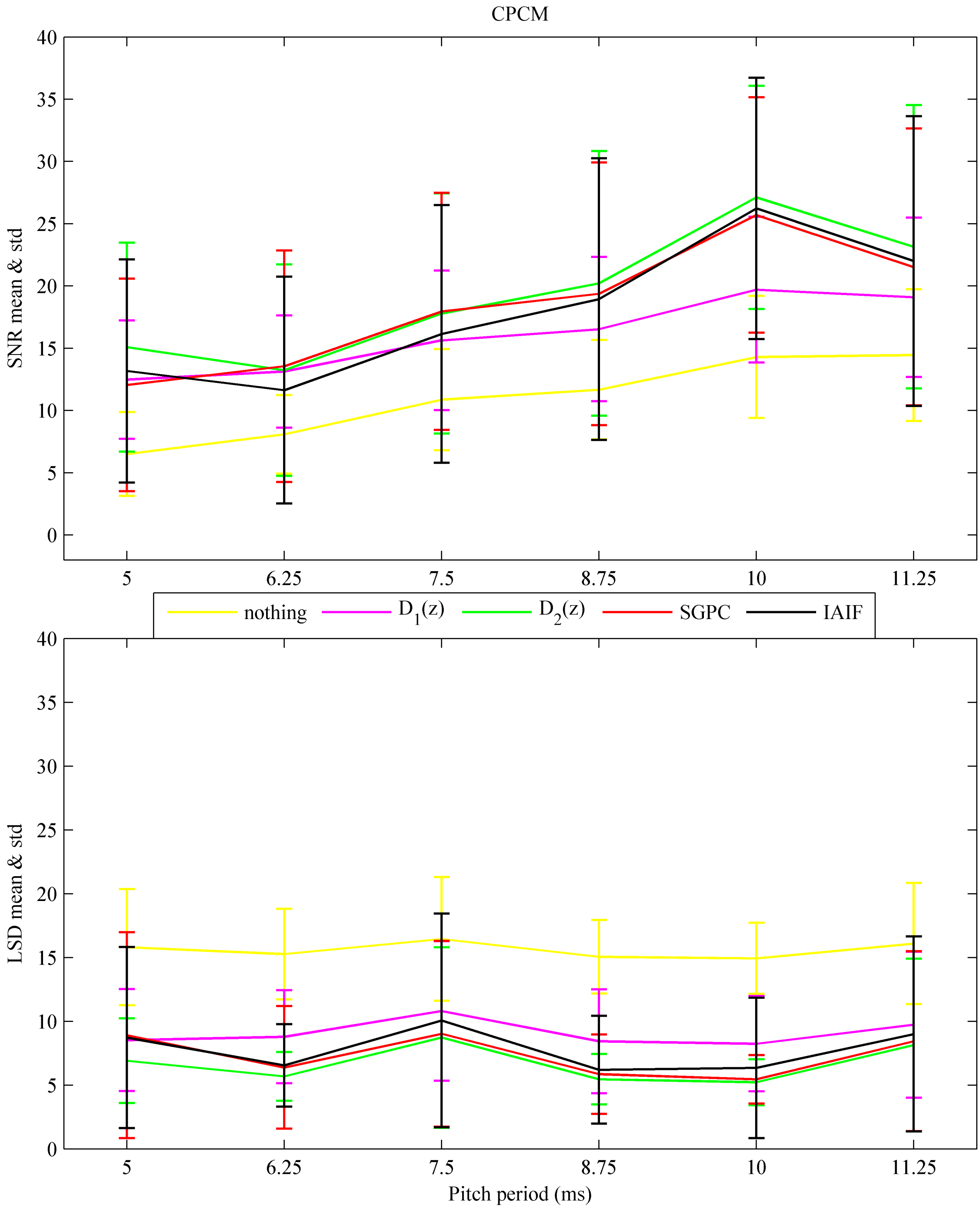


Figure 4-6: CPCM, (analysis interval= $p + 20$): means and standard deviations of SNRs and LSDs.

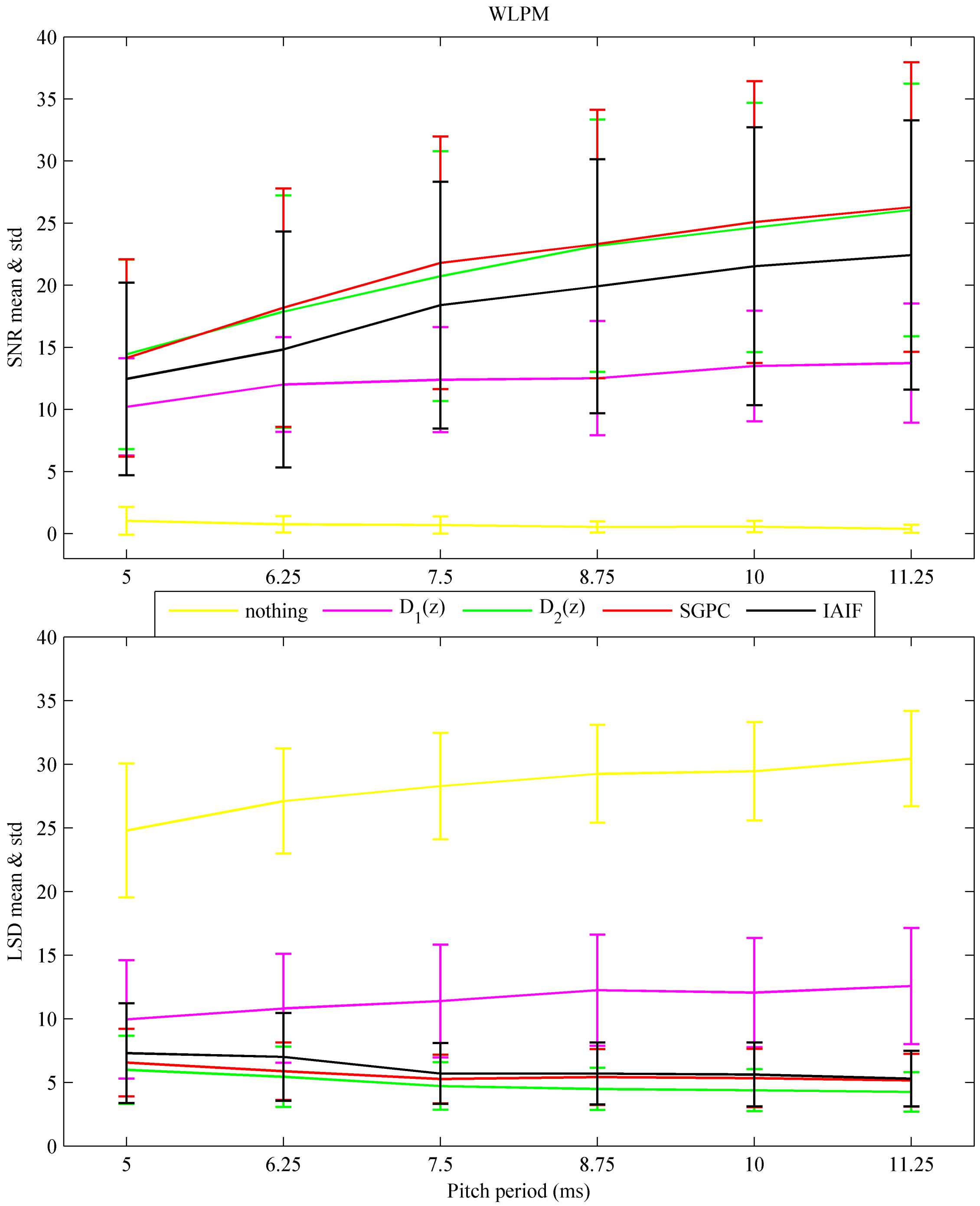


Figure 4-7: WLPM, $(M = 10, k = 1)$: means and standard deviations of SNRs and LSDs.

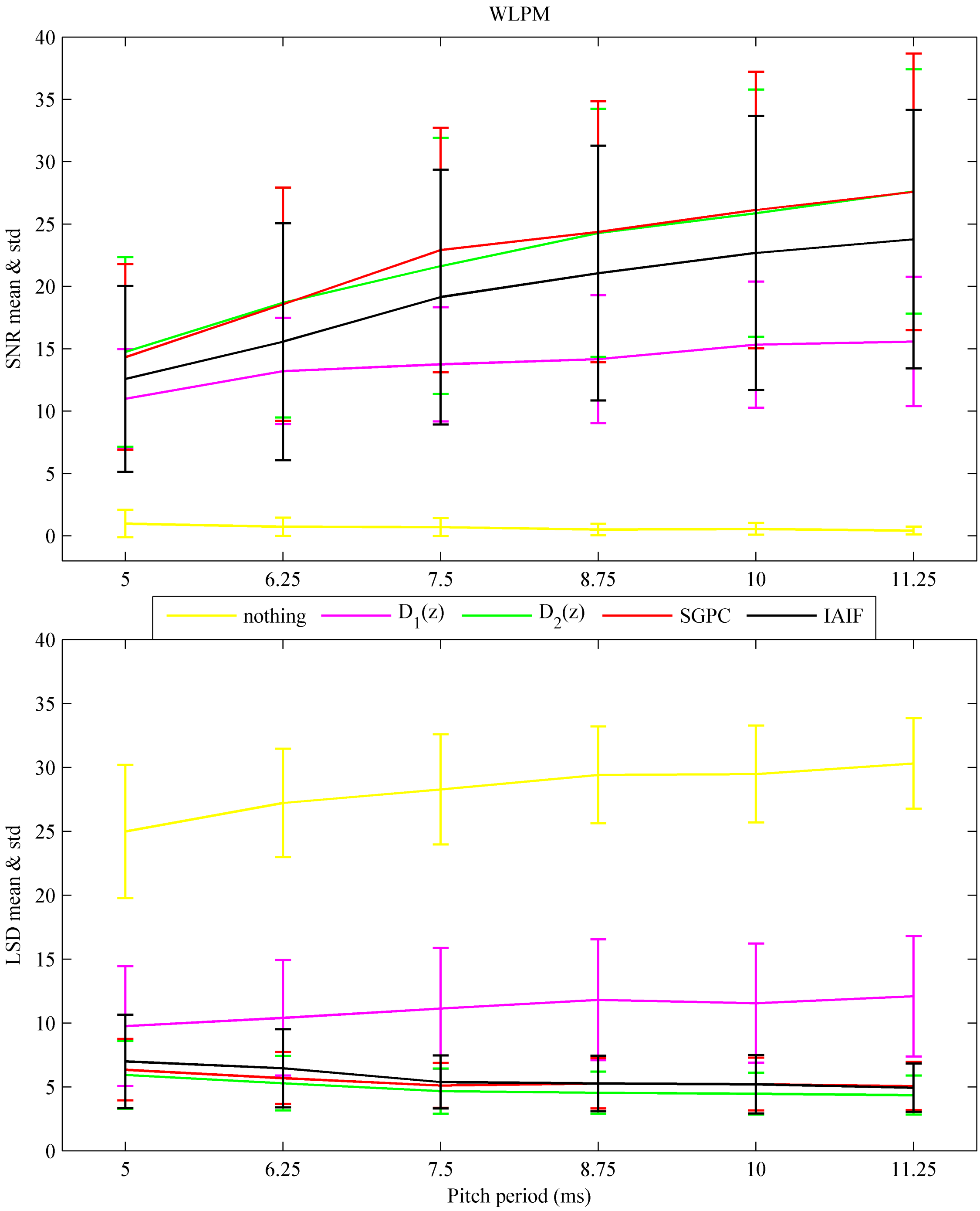


Figure 4-8: WLPM, $(M = 15, k = 1)$: means and standard deviations of SNRs and LSDs.

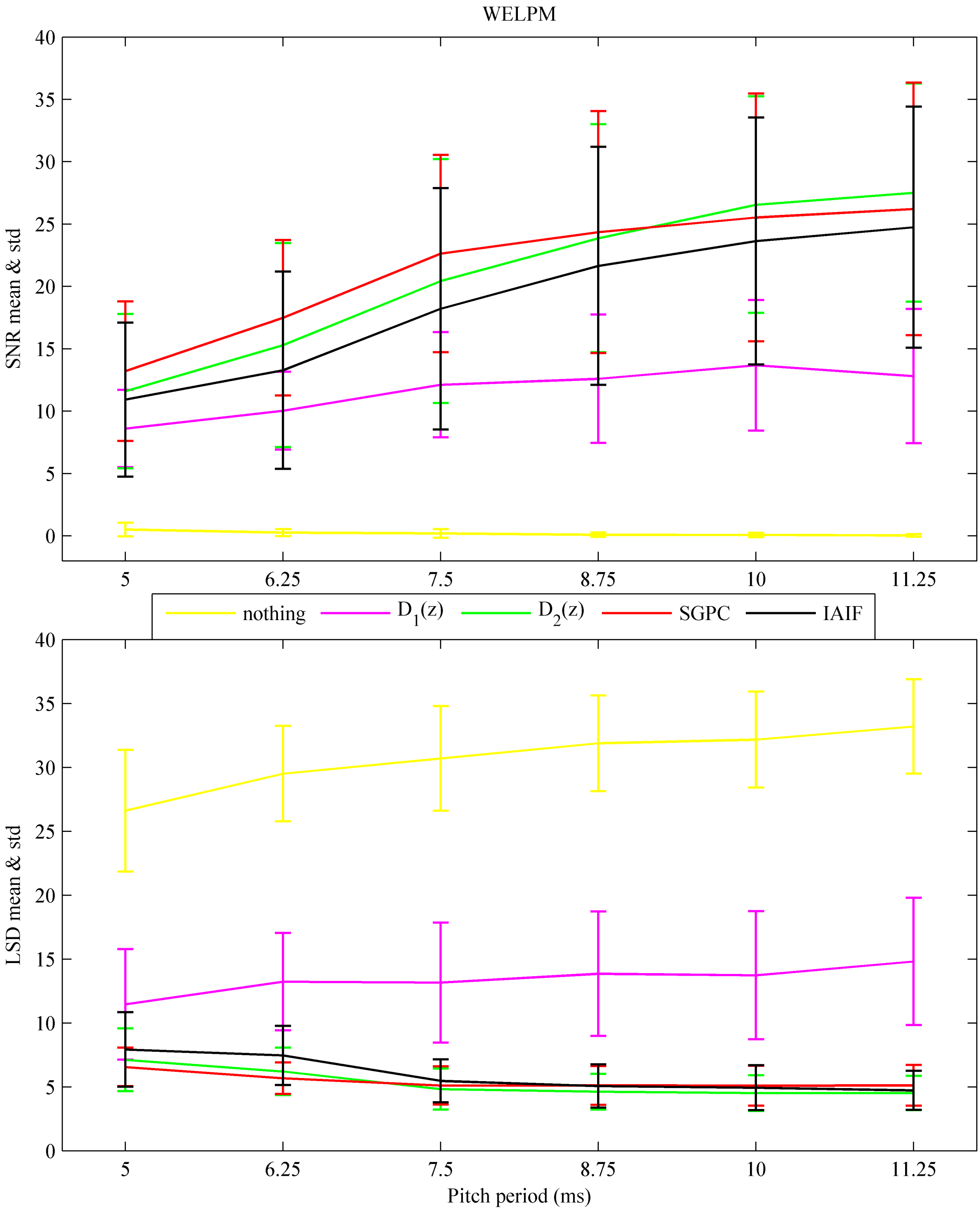


Figure 4-9: WELPM, $(\sigma = 31, \kappa = 0.89)$: means and standard deviations of SNRs and LSDs.

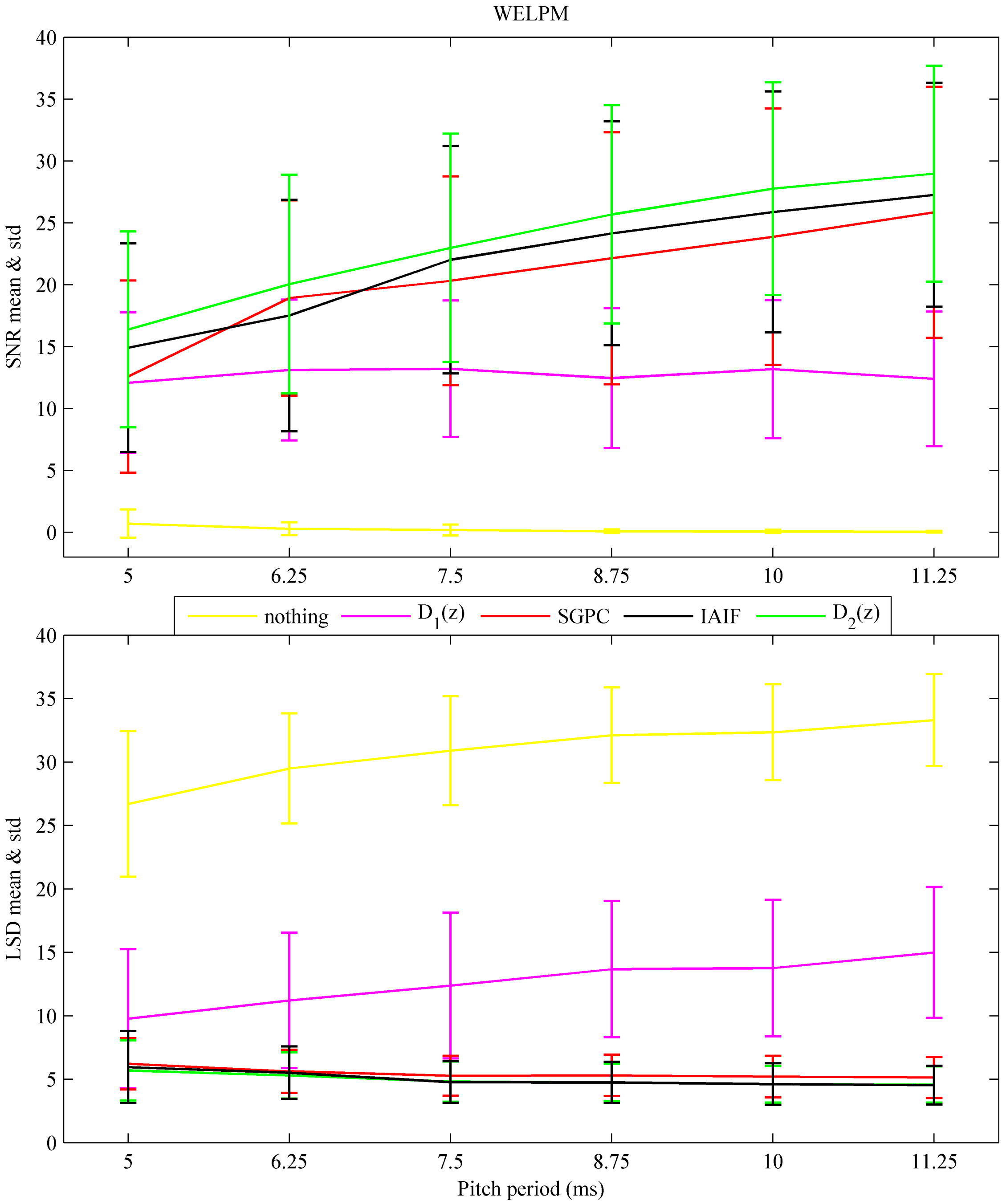


Figure 4-10: WELPM, $(\sigma = 10, \kappa = 0.89)$: means and standard deviations of SNRs and LSDs.

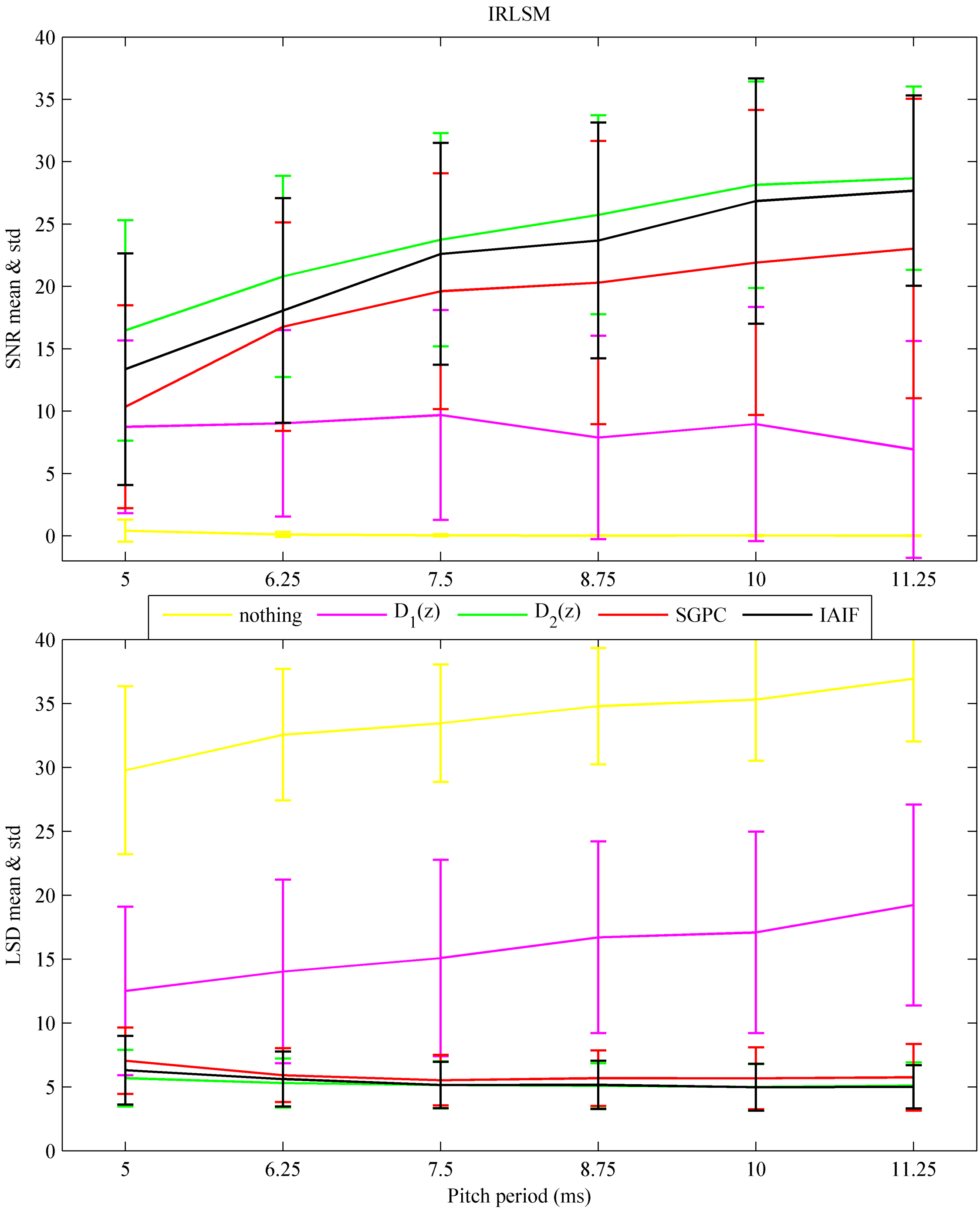


Figure 4-11: IRLSM: means and standard deviations of SNRs and LSDs.

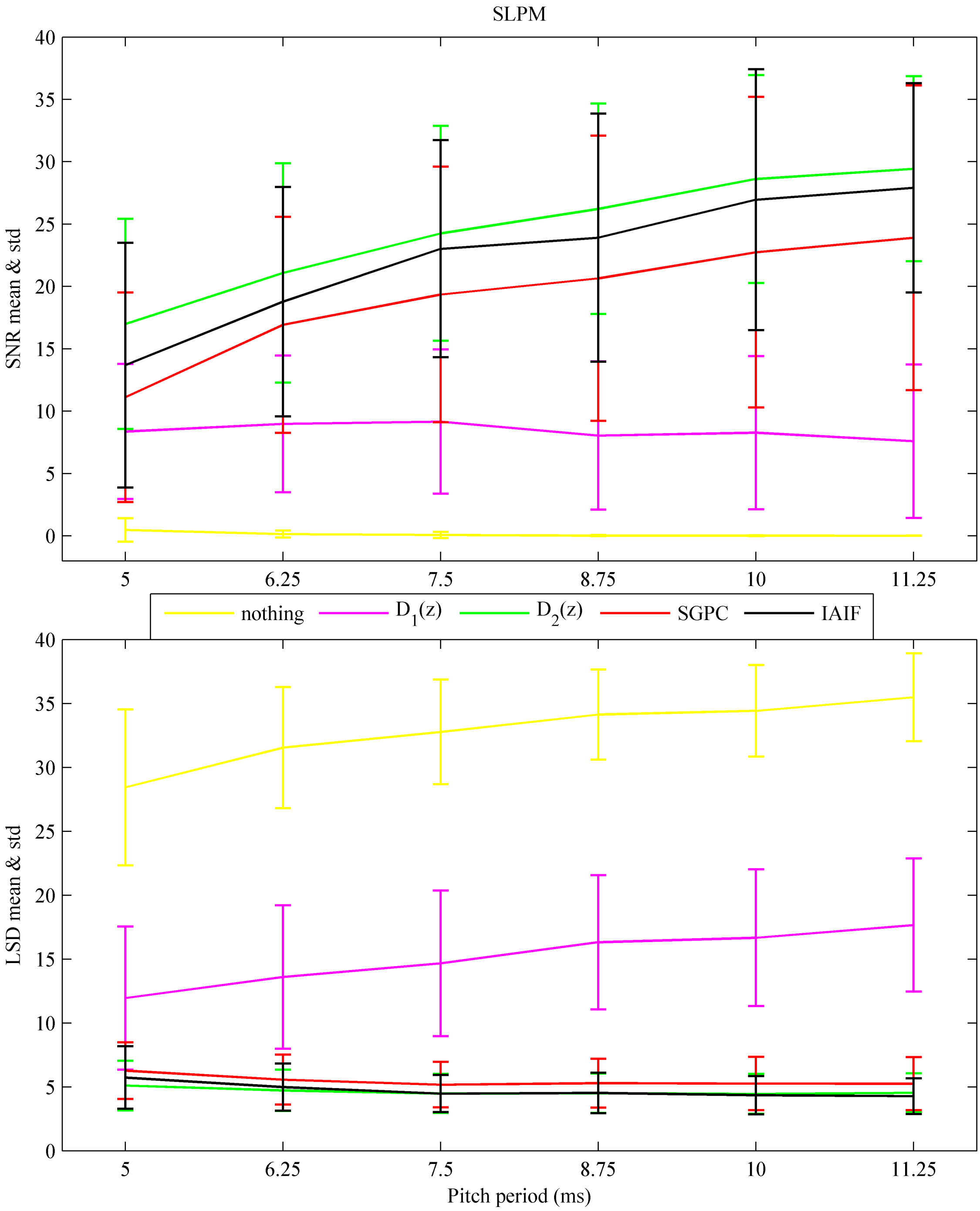


Figure 4-12: SLPM: means and standard deviations of SNRs and LSDs.

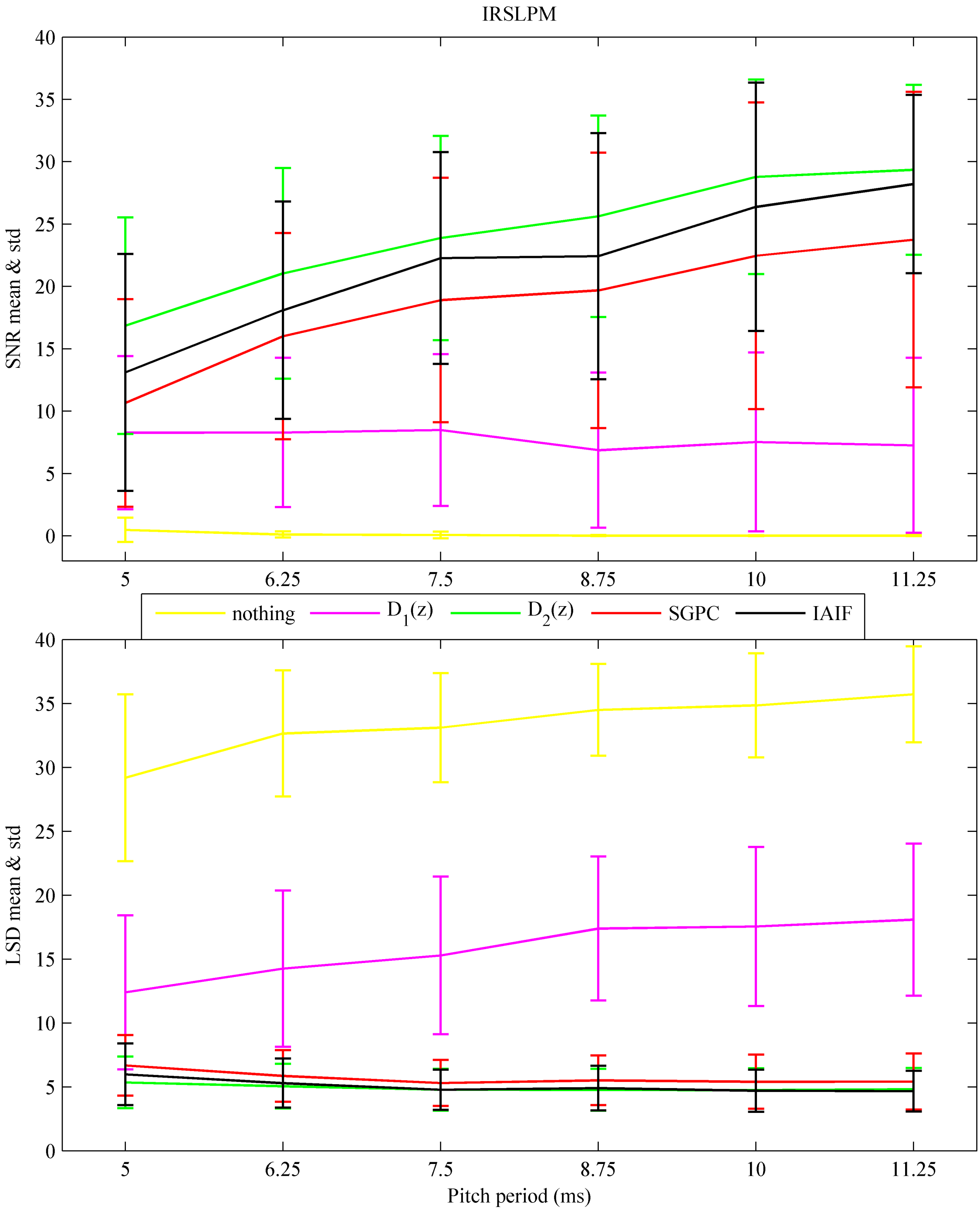


Figure 4-13: IRSLPM: means and standard deviations of SNRs and LSDs.

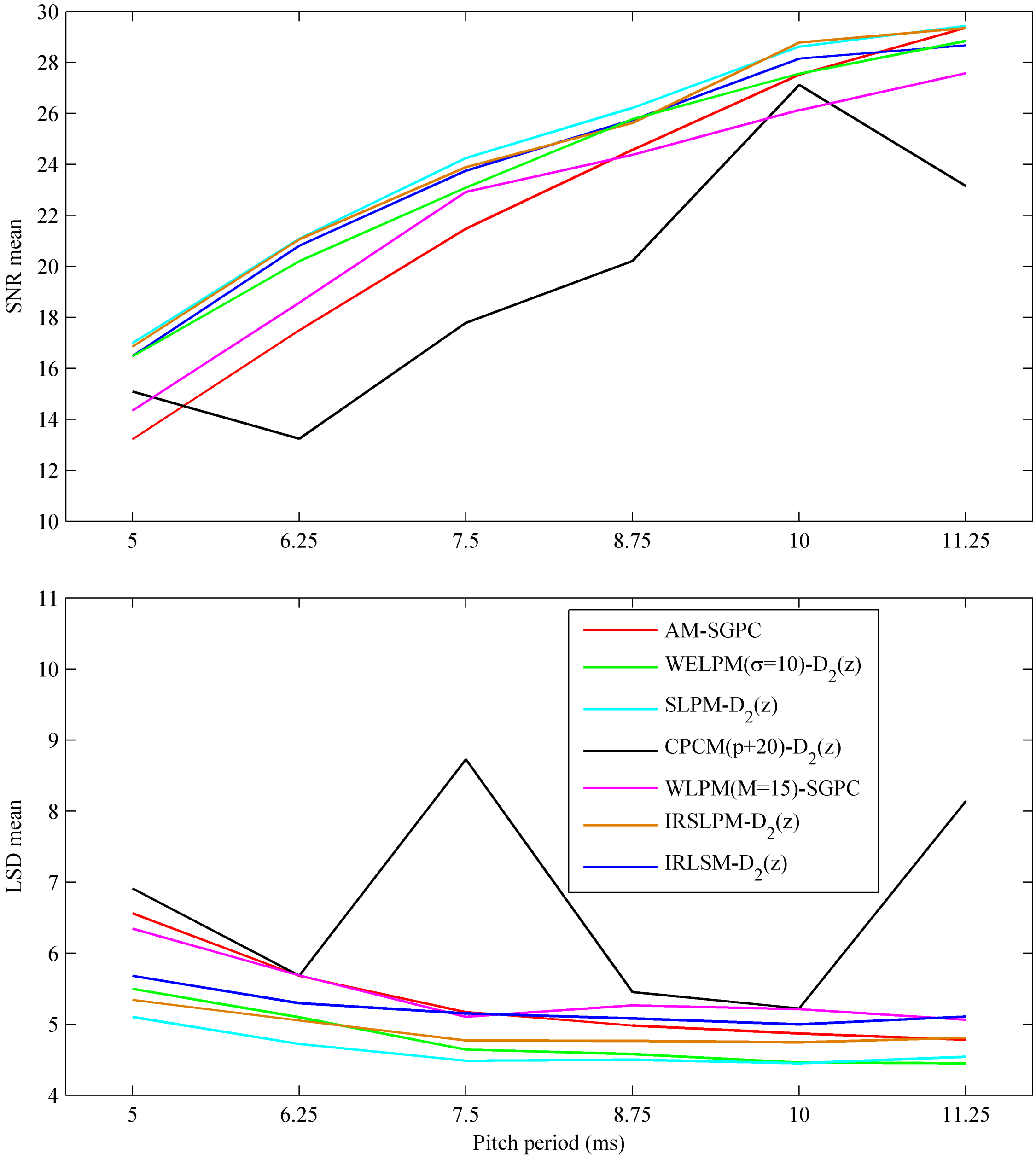


Figure 4-14: Best combinations: means of SNRs and LSDs.

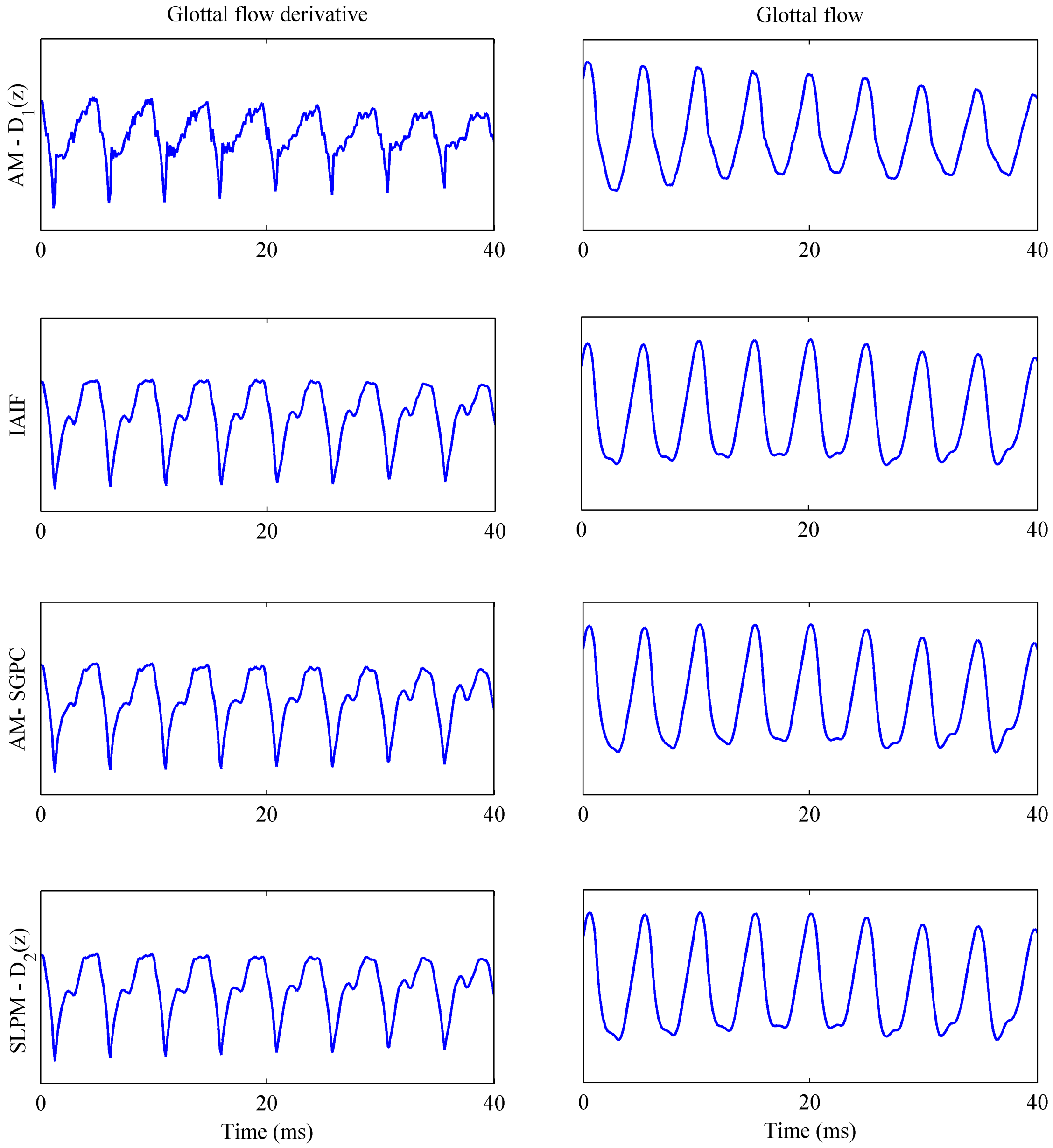


Figure 4-15: Estimated glottal flows and glottal flow derivatives of a real speech signal using various combinations.

Reverberated speech: Now we repeat the previous experiments on the same synthetic signals under constant reverberation with $T_{60} = 0.4$ s. We test only the best combinations of Figure 4-14. As we can see in the top sub-figure of Figure 4-16, an accurate estimate of the glottal flow derivative from reverberated speech cannot be obtained. However, if we apply dereverberation prior to inverse filtering this may allow us to estimate the glottal flow derivative. The results of the bottom sub-figure of Figure 4-16 state that we can obtain a quite accurate estimate of the true vocal tract for high pitch period speakers and especially with the combinations (WLPM, $D_2(z)$) and (SLPM, $D_2(z)$). Since we can estimate the vocal tract quite accurately from the reverberated speech, if we enhance in somehow the reverberated residual, we can combine these two in order to obtain a deverberated speech signal. Section 4-3 explores this application.

An other important factor which will possibly improve the performance of a such dereverberation method is to obtain the sparsest possible residuals. Thus, in the next section several combinations are evaluated with and without reverberation in terms of sparsity and stability.

4-2-2 Sparsity, Stability and Robustness to Reverberation

We use 10 speech recordings, produced by 5 males and 5 females, from the APLAWD database [87]. Each recording consists of a voiced sentence and two breathy parts at the start and at the end. These two breathy parts are not considered to our experiments. Nevertheless, there are a few frames inside the voiced sentences which are breathy and some are semi-vowels (i.e., /l/, /w/, /y/) [4]. Therefore, in total, we use 796 speech frames of length 40 ms. The sampling frequency is $f_s = 8$ KHz and the tested LP orders are $q = 10, 16$. WELPM is evaluated using $\sigma = 10$ and $\kappa = 0.89$, while WLPM is evaluated using $M = 15$. Note that the parametrization of these two methods might not be optimal. We selected these parameterizations according to our observations in Subsection 4-2-1.

All L_1 optimization problems are solved using the primal-dual interior point algorithm of the l1-magic toolbox [88]. For consistency, the average Gini values for each combination are computed over the maximum possible intersected set of frames that are stable for all combinations. The only method that is not included is CPCM due to its high probability of obtaining non-stable frames for reverberated and non-reverberated speech, i.e., approximately 75% and 35%, respectively. The experiments are carried out in a reverberant and a non-reverberant environment in order to test the robustness of these methods in terms of stability and sparsity. The source-image method [89, 90] is used for simulating a reverberant room with dimensions 6x5x4 m. The reverberation time T_{60} varies between 0.2 s and 0.8 s with steps of 0.2 s. Five different talker-microphone position pairs are used for the evaluation and they are placed randomly in the inner concentric room box of 5x4x3 m.

Tables 4-1 and 4-2 show the average Gini values per frame and probabilities of stability (i.e., the ratio (number of stable frames)/(total number of frames=796)) for each combination using as the LP order $q = 10$. Tables 4-3 and 4-4 show the same results for $q = 16$. In this section, a frame is considered unstable when its corresponding estimated filter has at least one pole outside the unit circle. It is clear from Tables 4-1 and 4-3 that the sparsest methods are IRLSM and IRSLPM. Moreover, note that in case of $q = 10$, almost always, pre-emphasis or glottal-cancellation *increases* the sparsity of the residual. The same happens for $q = 16$ but there are are some rare cases in which pre-emphasis or glottal-cancellation reduces sparsity

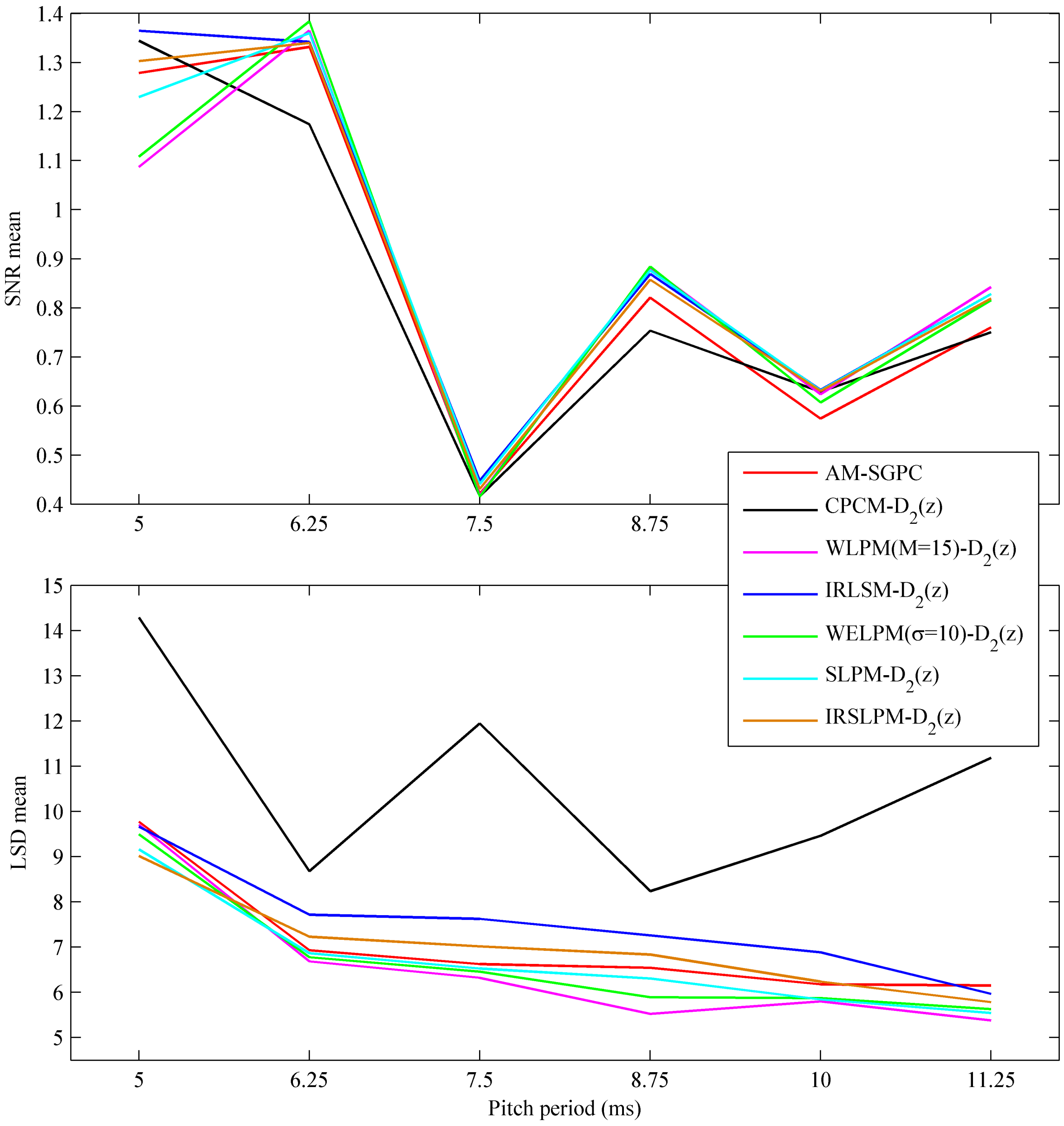


Figure 4-16: Best combinations for reverberated speech: means of SNRs and LSDs.

(see the explanation in Subsection 3-2-2). It should be mentioned that IRSLPM is not much sparser than SLPM. Furthermore, as was expected, when there is no pre-emphasis or glottal-cancellation prior to speech analysis, an increment of order means a sparser residual (see the explanation in Section 3-1).

Pre-emphasis and glottal-cancellation *increases* the percentage of stable filters. Furthermore, note that by increasing the LP order the stability decreases. This happens because when the LP order is higher than the true order of the vocal tract, sometimes the poles of the maximum-phase component of the glottal flow are estimated as well. This may not be a bad behavior in some applications. For instance the high instability of IRLSM means that

Combination \ T_{60}	0	0.2	0.4	0.6	0.8
(AM, nothing)	0.4939	0.4423	0.4213	0.4213	0.4197
(AM, $D_1(z)$)	0.4998	0.4452	0.4262	0.4223	0.4204
(AM, $D_2(z)$)	0.5127	0.4468	0.4266	0.4228	0.4213
(AM, IAIFGC)	0.5112	0.4459	0.4267	0.4231	0.4214
(AM, SGPC)	0.5002	0.4437	0.4260	0.4215	0.4205
(WLPM, nothing)	0.4986	0.4443	0.4270	0.4229	0.4215
(WLPM, $D_1(z)$)	0.5064	0.4481	0.4276	0.4231	0.4214
(WLPM, $D_2(z)$)	0.5233	0.4496	0.4278	0.4238	0.4221
(WLPM, IAIFGC)	0.5212	0.4485	0.4277	0.4240	0.4224
(WLPM, SGPC)	0.5066	0.4458	0.4269	0.4222	0.4212
(WELPM, nothing)	0.4843	0.4467	0.4261	0.4217	0.4202
(WELPM, $D_1(z)$)	0.4941	0.4494	0.4267	0.4222	0.4206
(WELPM, $D_2(z)$)	0.5061	0.4508	0.4268	0.4232	0.4217
(WELPM, IAIFGC)	0.5047	0.4495	0.4271	0.4234	0.4214
(WELPM, SGPC)	0.4957	0.4470	0.4263	0.4217	0.4209
(IRLSM, nothing)	0.5573	0.4770	0.4555	0.4512	0.4486
(IRLSM, $D_1(z)$)	0.5601	0.4802	0.4559	0.4520	0.4495
(IRLSM, $D_2(z)$)	**0.5666**	0.4811	0.4564	0.4510	0.4498
(IRLSM, IAIFGC)	0.5653	0.4797	0.4562	0.4514	0.4502
(IRLSM, SGPC)	0.5591	0.4779	0.4559	0.4515	0.4498
(SLPM, nothing)	0.5419	0.4698	0.4503	0.4460	0.4444
(SLPM, $D_1(z)$)	0.5434	0.4731	0.4512	0.4472	0.4449
(SLPM, $D_2(z)$)	0.5523	0.4741	0.4509	0.4467	0.4456
(SLPM, IAIFGC)	0.5511	0.4728	0.4511	0.4472	0.4455
(SLPM, SGPC)	0.5428	0.4707	0.4505	0.4463	0.4451
(IRSLPM, nothing)	0.5509	0.4733	0.4557	0.4533	0.4522
(IRSLPM, $D_1(z)$)	0.5546	0.4809	**0.4593**	0.4556	0.4536
(IRSLPM, $D_2(z)$)	0.5619	**0.4820**	**0.4593**	0.4551	0.4540
(IRSLPM, IAIFGC)	0.5609	0.4804	**0.4593**	**0.4561**	**0.4541**
(IRSLPM, SGPC)	0.5539	0.4781	0.4586	0.4544	0.4539

Table 4-1: Average Gini values for various reverberation times T_{60}. The LP order is $q = 10$. Bold numbers indicate the largest Gini value per column.

this method may estimate the maximum-phase component of the glottal flow more accurately. Thus, we recommend a future research on this idea. It is worth noting that reverberation does *not* decrease stability. On the contrary, in some occasions it might increase the percentage of stability especially when we are using a higher LP order than the true order of the vocal tract.

Figure 4-17 depicts a speech frame without reverberation and the corresponding residuals of some combinations for two different prediction orders $q = 10$ and 16. It is clear that the least sparse combinations are of the method AM. The differences in sparsity of all the other combinations are difficult to be observed. The parts that contribute mostly in an

T_{60} Combination	0	0.2	0.4	0.6	0.8
(AM, nothing)	1	1	1	1	1
(AM, $D_1(z)$)	1	1	1	1	1
(AM, $D_2(z)$)	1	1	1	1	1
(AM, IAIFGC)	1	1	1	1	1
(AM, SGPC)	1	1	1	1	1
(WLPM, nothing)	0.9987	0.9955	0.9980	0.9985	0.9992
(WLPM, $D_1(z)$)	0.9987	**0.9987**	**0.9997**	**1**	**1**
(WLPM, $D_2(z)$)	0.9975	0.9967	0.9992	**1**	**1**
(WLPM, IAIFGC)	0.9975	0.9980	0.9992	**1**	**1**
(WLPM, SGPC)	**1**	0.9980	0.9990	**1**	**1**
(WELPM, nothing)	0.9724	0.9882	0.9897	0.9950	0.9940
(WELPM, $D_1(z)$)	0.9761	0.9925	0.9920	0.9962	0.9965
(WELPM, $D_2(z)$)	0.9736	0.9942	0.9977	0.9972	0.9987
(WELPM, IAIFGC)	0.9711	0.9950	0.9970	0.9972	0.9987
(WELPM, SGPC)	0.9812	0.9937	0.9960	0.9982	0.9987
(IRLSM, nothing)	0.8505	0.9030	0.9445	0.9583	0.9608
(IRLSM, $D_1(z)$)	0.9510	0.9616	0.9603	0.9683	0.9714
(IRLSM, $D_2(z)$)	0.9912	0.9709	0.9668	0.9716	0.9734
(IRLSM, IAIFGC)	0.9912	0.9638	0.9673	0.9696	0.9721
(IRLSM, SGPC)	0.9799	0.9590	0.9641	0.9709	0.9799
(SLPM, nothing)	0.9636	0.9864	0.9925	0.9962	0.9962
(SLPM, $D_1(z)$)	**1**	0.9957	0.9977	0.9972	0.9977
(SLPM, $D_2(z)$)	**1**	0.9980	0.9985	0.9977	0.9977
(SLPM, IAIFGC)	**1**	0.9967	0.9982	0.9992	0.9982
(SLPM, SGPC)	0.9987	0.9962	0.9962	0.9975	0.9985
(IRSLPM, nothing)	0.9384	0.9545	0.9698	0.9839	0.9864
(IRSLPM, $D_1(z)$)	0.9912	0.9872	0.9892	0.9917	0.9910
(IRSLPM, $D_2(z)$)	0.9987	0.9915	0.9889	0.9894	0.9902
(IRSLPM, IAIFGC	0.9975	0.9884	0.9884	0.9897	0.9899
(IRSLPM, SGPC)	0.9962	0.9872	0.9894	0.9920	0.9902

Table 4-2: Probabilities of Stability for different reverberation times T_{60}. The LP order is $q = 10$. Bold numbers indicate the largest probability of stability per column (not including the AMmethod).

increased sparsity are the amplitudes of the main epochs. For instance, observe that the epochs of (IRLSM, $D_2(z)$) are much stronger than the corresponding of the combinations of AM. Another interesting observation is that AM *cannot* estimate the zero from the lips while the other methods estimate it more accurately. This can be show by the negative epochs which are one sample ahead the positive epochs. The increased sparsity and the more accurate estimation of the lips are the factors that should be considered in Section 4-3 in order to select the proper analysis method for dereverberation. Finally, note that the increased order, increase the sparsity as was expected. This can be viewed better in Figure 4-18 in which a reverberated speech signal is analyzed. Furthermore, note in this figure, that there

T_{60} Combination	0	0.2	0.4	0.6	0.8
(AM, nothing)	0.5029	0.4440	0.4255	0.4221	0.4202
(AM, $D_1(z)$)	0.5051	0.4450	0.4262	0.4219	0.4203
(AM, $D_2(z)$)	0.5124	0.4459	0.4262	0.4221	0.4205
(AM, IAIFGC)	0.5110	0.4453	0.4260	0.4221	0.4208
(AM, SGPC)	0.5047	0.4438	0.4254	0.4225	0.4213
(WLPM, nothing)	0.5100	0.4456	0.4283	0.4248	0.4226
(WLPM, $D_1(z)$)	0.5165	0.4488	0.4289	0.4245	0.4223
(WLPM, $D_2(z)$)	0.5228	0.4495	0.4281	0.4242	0.4221
(WLPM, IAIFGC)	0.5218	0.4490	0.4282	0.4245	0.4224
(WLPM, SGPC)	0.5143	0.4471	0.4274	0.4243	0.4231
(WELPM, nothing)	0.4942	0.4503	0.4289	0.4243	0.4224
(WELPM, $D_1(z)$)	0.4982	0.4514	0.4295	0.4243	0.4223
(WELPM, $D_2(z)$)	0.5046	0.4522	0.4289	0.4243	0.4227
(WELPM, IAIFGC)	0.5032	0.4517	0.4291	0.4245	0.4227
(WELPM, SGPC)	0.4987	0.4502	0.4283	0.4245	0.4231
(IRLSM, nothing)	0.5796	0.4964	0.4726	0.4692	0.4668
(IRLSM, $D_1(z)$)	0.5802	0.4983	0.4734	0.4694	0.4672
(IRLSM, $D_2(z)$)	**0.5828**	**0.4986**	0.4731	0.4695	0.4663
(IRLSM, IAIFGC)	0.5824	0.4974	0.4734	0.4698	0.4666
(IRLSM, SGPC)	0.5784	0.4960	0.4732	0.4696	0.4675
(SLPM, nothing)	0.5618	0.4841	0.4630	0.4596	0.4574
(SLPM, $D_1(z)$)	0.5625	0.4861	0.4635	0.4597	0.4575
(SLPM, $D_2(z)$)	0.5647	0.4863	0.4635	0.4597	0.4574
(SLPM, IAIFGC)	0.5642	0.4855	0.4639	0.4599	0.4579
(SLPM, SGPC)	0.5609	0.4838	0.4633	0.4597	0.4578
(IRSLPM, nothing)	0.5727	0.4899	0.4722	0.4697	0.4687
(IRSLPM, $D_1(z)$)	0.5750	0.4969	0.4753	0.4719	0.4696
(IRSLPM, $D_2(z)$)	0.5775	0.4975	0.4753	0.4716	0.4695
(IRSLPM, IAIFGC)	0.5770	0.4962	**0.4755**	**0.4720**	**0.4698**
(IRSLPM, SGPC)	0.5738	0.4943	0.4752	0.4721	0.4696

Table 4-3: Average Gini values for various reverberation times T_{60}. The LP order is $q = 16$. Bold numbers indicate the largest Gini value per column.

are additional epochs to the residuals due to the reflective surfaces. This extra contribution decreases the sparsity of the reverberated residuals compared to the clean ones.

T_{60} Combination	0	0.2	0.4	0.6	0.8
(AM, nothing)	1	1	1	1	1
(AM, $D_1(z)$)	1	1	1	1	1
(AM, $D_2(z)$)	1	1	1	1	1
(AM, IAIFGC)	1	1	1	1	1
(AM, SGPC)	1	1	1	1	1
(WLPM, nothing)	0.9786	0.9701	0.9844	0.9867	0.9897
(WLPM, $D_1(z)$)	0.9962	**0.9920**	**0.9972**	**0.9975**	0.9982
(WLPM, $D_2(z)$)	0.9849	0.9864	0.9942	0.9972	**0.9985**
(WLPM, IAIFGC)	0.9824	0.9869	0.9927	0.9970	0.9982
(WLPM, SGPC)	0.9962	0.9859	0.9932	0.9970	0.9982
(WELPM, nothing)	0.9183	0.9656	0.9711	0.9796	0.9827
(WELPM, $D_1(z)$)	0.9246	0.9779	0.9814	0.9925	0.9925
(WELPM, $D_2(z)$)	0.9095	0.9771	0.9814	0.9887	0.9889
(WELPM, IAIFGC)	0.9158	0.9769	0.9819	0.9905	0.9902
(WELPM, SGPC)	0.9221	0.9701	0.9711	0.9847	0.9849
(IRLSM, nothing)	0.6307	0.8000	0.8769	0.8807	0.8945
(IRLSM, $D_1(z)$)	0.8907	0.8673	0.8937	0.9030	0.9103
(IRLSM, $D_2(z)$)	0.9711	0.8824	0.8905	0.8987	0.9075
(IRLSM, IAIFGC)	0.9661	0.8789	0.8920	0.8982	0.9098
(IRLSM, SGPC)	0.9447	0.8693	0.8905	0.9010	0.9088
(SLPM, nothing)	0.8568	0.9603	0.9774	0.9839	0.9894
(SLPM, $D_1(z)$)	0.9849	0.9807	0.9884	0.9915	0.9942
(SLPM, $D_2(z)$)	**1**	0.9837	0.9879	0.9887	0.9907
(SLPM, IAIFGC)	**1**	0.9827	0.9874	0.9917	0.9912
(SLPM, SGPC)	0.9987	0.9789	0.9849	0.9894	0.9897
(IRSLPM, nothing)	0.7726	0.8804	0.9327	0.9530	0.9618
(IRSLPM, $D_1(z)$)	0.9560	0.9573	0.9671	0.9724	0.9771
(IRSLPM, $D_2(z)$)	0.9550	0.9588	0.9673	0.9636	0.9688
(IRSLPM, IAIFGC	0.9912	0.9563	0.9638	0.9701	0.9761
(IRSLPM, SGPC)	0.9925	0.9533	0.9606	0.9671	0.9678

Table 4-4: Probabilities of Stability for different reverberation times T_{60}. The LP order is $q = 16$. Bold numbers indicate the largest probability of stability per column (not including the AM method).

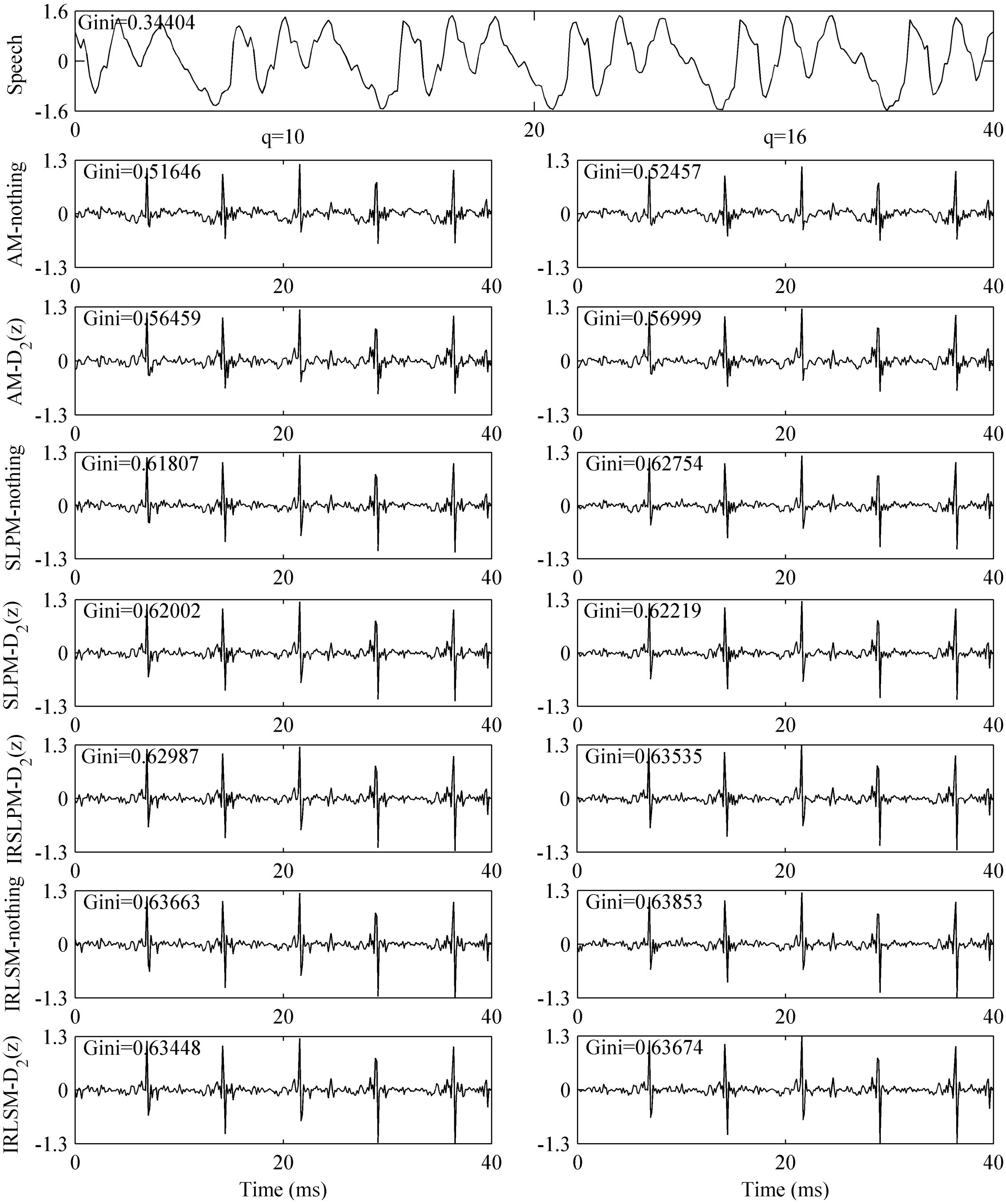


Figure 4-17: Residual of various combinations for two different LP orders, $q = 10, 16$.

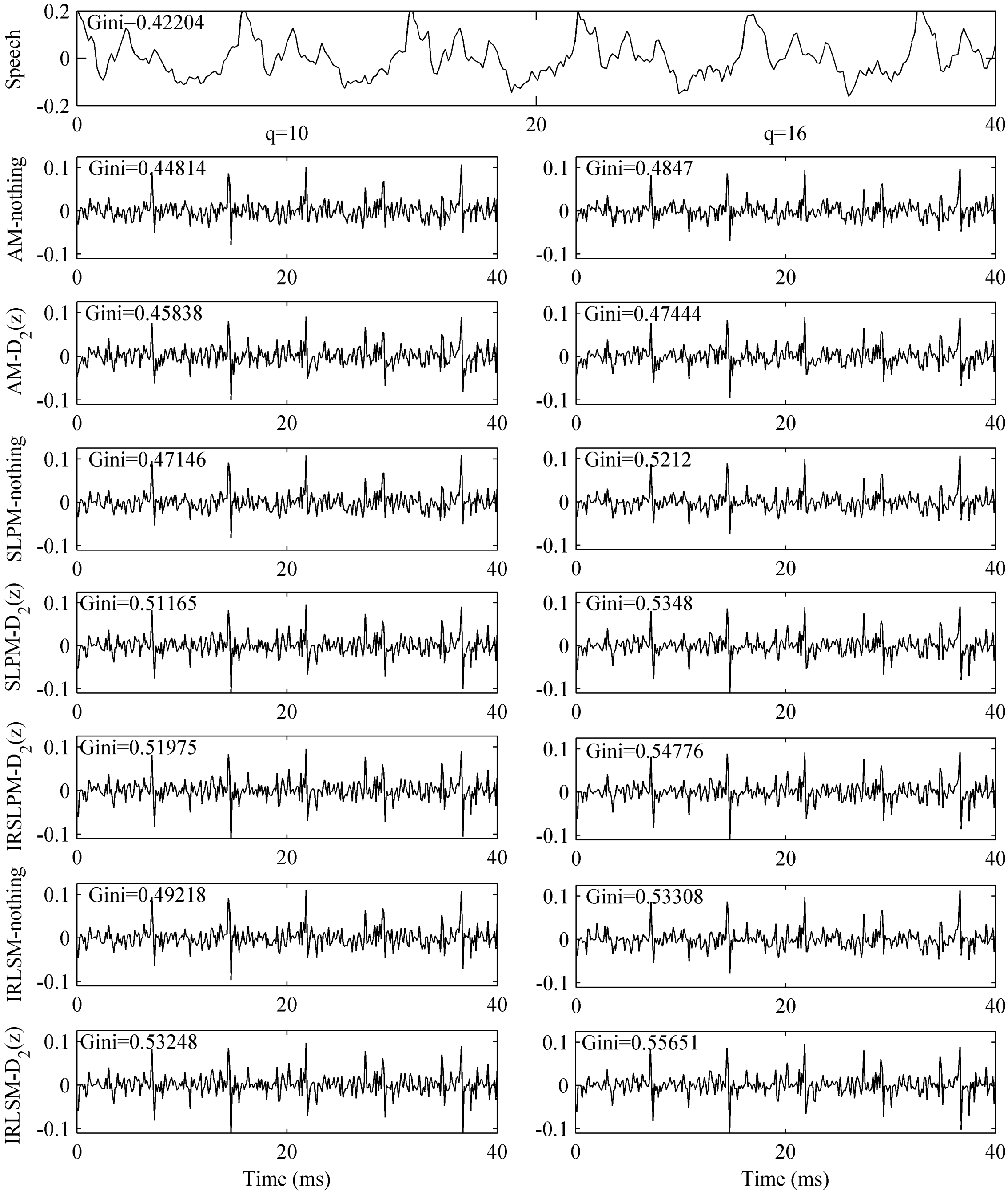


Figure 4-18: Reverberated ($T_{60} = 0.6$ s) residual of various combinations for two different LP orders, $q = 10, 16$.

4-3 Speech Dereverberation

In this section, we explore the performance of several combinations in the context of the dereverberation application proposed in [1,2]. In particular, the speech signal $x[n]$ is acquired via a small number, M, of microphones placed in different positions, i.e., (x_i, y_i, z_i), for $i = 1, 2, ..., M$. The acquired speech signals $x_i[n]$ are given by

$$x_i[n] = h_i[n] * x[n], \tag{4-6}$$

where $h_i[n]$ is the room impulse response from the position of the source to the position of the i^{th} microphone. We know from Chapter 2 that according to SFM, the speech signal can be written as

$$x[n] = h[n] * e[n], \tag{4-7}$$

where $h[n]$ is the impulse response of the filter and $e[n]$ is the source signal or the prediction error sequence. If we replace $x[n]$ in Equation 4-6 with Equation 4-7, we obtain

$$x_i[n] = h_i[n] * (h[n] * e[n]) = \bar{h}_i[n] * e_i[n], \tag{4-8}$$

Therefore, now the LPCs of Equation 3-2 are different and are denoted as b_{ki} instead of a_k. It was empirically shown in [2] that

$$\bar{a}_k = \frac{1}{M}\sum_{i=1}^{M} b_{ki} \approx a_k. \tag{4-9}$$

Moreover, the authors of [1] after aligning the residuals $e_i[n]$, they calculated the Hilbert envelopes of them as

$$\hat{e}_i[n] = \sqrt{e_i^2[n] + e_{iH}^2[n]}, \tag{4-10}$$

where $e_{iH}[n]$ is the Hilbert transform of $e_i[n]$. Then, they combined all the obtained Hilbert envelopes as

$$\hat{e}_c[n] = \sqrt{\sum_{i=1}^{M} \hat{e}_i^2[n]}. \tag{4-11}$$

They claimed that they weighted the residual $e_1[n]$ (e.g. the residual of the closest microphone) as follows

$$e_{1c}[n] = \frac{\sum_n e_1[n]\hat{e}_c[n]}{\sum_n \hat{e}_c[n]}, \tag{4-12}$$

in order to obtain the enchanced residual $e_{1c}[n]$. However, this operation produces a constant value for every n of the signal $e_{1c}[n]$. So, to the author's opinion there is a typo in this formula and this typo is the summation of the numerator. Therefore, the correct formula is

$$e_{1c}[n] = \frac{e_1[n]\hat{e}_c[n]}{\sum_n \hat{e}_c[n]}. \tag{4-13}$$

We observed that, if in the numerator we replace $e_1[n]$ with the average of all the aligned residuals, the weighting scheme behaves better. Thus, our formula is the following

$$e_{1c}[n] = \frac{\frac{1}{M}\sum_{i=1}^{M}(e_i[n])\hat{e}_c[n]}{\sum_n \hat{e}_c[n]}. \tag{4-14}$$

This weighting procedure emphasizes the true epochs of the residual and de-emphasizes all the other values including the secondary peaks caused by reverberation. Therefore, once the residual of the first microphone is enhanced, it can be combined with the filter

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} \bar{a}_k z^{-k}} \tag{4-15}$$

producing the dereverberated speech signal $x_c[n]$. Note that if we have applied a pre-emphasis filter prior to LP analysis, we should also de-emphasize the signal at the end. It is worth noting, that this is not an easy procedure when some samples of the residuals are modified. In our future work we will try to find an efficient de-emphasis method in this specific case.

The residuals in both papers [1, 2] are obtained with the combination (AM, nothing). Here we will obtain the residuals with three different combinations: (AM, nothing), (WLPM, $D_2(z)$) and (SLPM, $D_2(z)$). Figure 4-19 shows the residuals of the cleans speech signal, the reverberated speech signals acquired in three microphones, and the enhanced speech signal using two different methods. The first method is a simple averaging of all aligned reverberated residuals, and the second is Equation 4-14. As we can see, the combinations (SLPM, $D_2(z)$), (WLPM, $D_2(z)$) produce an enchanced residual which looks closer to the true residual. Note that for this experiment we used 3 different microphones positioned in a linear array configuration. The distances between them is 10 cm and the reverberation time is $T_{60} = 0.3$ s.

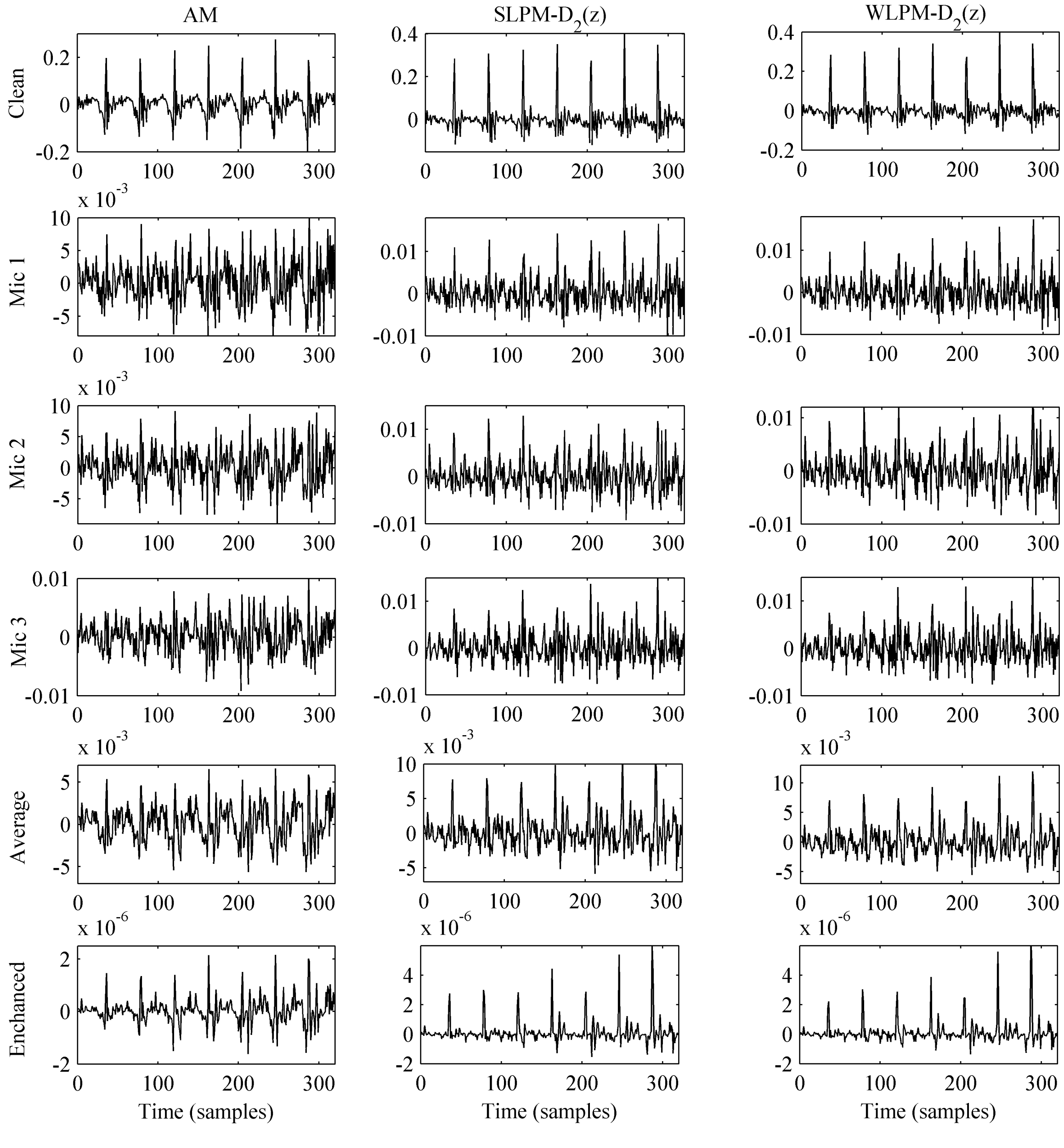


Figure 4-19: Residual enhancement using various combinations.

Chapter 5

Conclusion and Future Work

In the first part of the thesis we reviewed the speech production mechanism and its discrete-time/space realization, the source filter model (SFM). Several models of the glottal flow derivative waveform and of the vocal tract filter were presented. Furthermore, we reviewed seven linear prediction (LP) methods and their properties. We presented two pre-emphasis methods and two glottal-cancellation methods and their significance in the context of LP analysis.

The second part of the thesis explored the performance of the LP methods combined with the pre-emphasis and glottal-cancellation methods in the context of two general application areas. The first area consists of applications which aim to estimate the true glottal flow derivative signal. The second area consists of applications which aim to find a sparse residual.

5-1 Concluding Remarks

Our concluding remarks are summarized in the following six bullets.

1. The method proposed in [19], i.e., the combination (IRSLPM, $D_2(z)$), is better than IAIF in estimating the glottal flow derivative waveform. Moreover, although its non-iterative version, (SLPM, $D_2(z)$) produces slightly less sparse residuals, it performs slightly better in inverse filtering applications.
2. When the speech signal is subject to reverberation, the estimation of the true vocal tract is more accurate with the combinations (WLPM, $D_2(z)$) and (SLPM, $D_2(z)$). We showed also that these combinations are useful for speech dereverberation applications such as in [1, 2].
3. When the LP order is close to the true order of the vocal tract, i.e., $q = 10$, we obtained less than 7% of unstable filters for all combinations except of (IRLSM, nothing), On the other hand, when the order is higher, i.e., $q = 16$, we obtained less than 23% of unstable filters for all combinations except of the combinations of the IRLSM method. Moreover, pre-emphasis and glottal-cancellation increases the percentage of the stable filters. Generally the IRLSM method and especially, the combination (IRLSM, nothing) gives the highest percentage of unstable filters and, especially, for higher LP orders. Note

also that when the speech signal is subject to reverberation, in most cases the stability is increased slightly. This may happen, because the glottal formant (consisting of two poles outside the unit circle) cannot be estimated accurately anymore. We observed that in general, reverberation does not decrease stability.

4. CPCM has the best performance compared to all the other methods when we have an ideal closed phase region. However, this is not the case in most real speech signals. CPCM becomes more robust if it is combined with a pre-emphasis or glottal-cancellation technique, in the case of a longer analysis interval (in the direction of GOI) than the true closed phase region. It is worth noting that the estimation of GOIs is a difficult task and, therefore, we may want to estimate only the GCI location and take as the analysis interval a few samples after this GCI location. In the context of this idea, our experiments showed that indeed pre-emphasis or glottal-cancellation prior to CPCM improves performance. When we say GCI location we do not mean the t_e instants, but the t_c instants. Generally it is difficult to estimate these instants. Therefore, we proposed in the present thesis to use the instant of the local maximum of the speech signal after the t_e instant.
5. The sparsest methods are IRLSM, IRSLPM and SLPM. Moreover, note that in case of $q = 10$, almost always, pre-emphasis or glottal-cancellation increases the sparsity of the residual. The same happens for $q = 16$ but there are are some rare cases in which pre-emphasis or glottal-cancellation reduces sparsity. It should be mentioned that IRSLPM is not much sparser than SLPM. Furthermore, as was expected, when there is no pre-emphasis or glottal-cancellation prior to speech analysis, an increase of order means a sparser residual.
6. The combinations (SLPM, $D_2(z)$) and (WLPM, $D_2(z)$) are well suited in the context of the speech dereverberation method proposed in [1, 2]. We showed that the residual is enhanced more accurately by those two combinations than the combination (AM, nothing) that was used in [1, 2].

5-2 Future Work

The validation of the LP methods in terms of the estimation accuracy of the glottal flow derivative was undertaken using synthetic speech signals which do not include the source-filter non-linear interaction. We recommend for future investigation, to include some non-linear interaction and re-do some of the experiments and see if the performances are the same.

Furthermore, the high instability percentage of (IRLSM, nothing) may not be desired in speech coding applications, but it may be useful in inverse filtering applications. We observed that this high instability of IRLSM is caused from the estimated poles of the glottal formant (i.e., the maximum phase part of the glottal flow) when the order is increased more than the approximate true order of the vocal tract. We are planning in the future to exploit this property in order to have a better inverse filtering performance when the glottal flow formant is very close to the first formant of the vocal tract. Most methods like IAIF are problematic in these situations because they estimate both pairs of poles inside the unit circle and, therefore, they may cancel the first formant of the vocal tract and not the glottal formant. The method that we will propose in the future, first detects the poles of the glottal formant which are outside the unit circle and then cancels them. This method can be used also as alternative to the methods that try to estimate the glottal formant frequency [55, 91].

Finally, a validation of our proposed method for the estimation of the GCI positions would be useful. An idea which may improve our proposed method is the following. The singular value decomposition of the speech signal may help to detect the correct local maximum of the speech signal after a few samples of the detected t_e instant. This is because, sometimes, the speech signal does not have this ideal shape of Figure 2-10 and it may have several peaks close to the t_e instant. Thus, via the singular value decomposition, we can find a smoother speech signal, and to obtain a more accurate GCI instant.

Bibliography

[1] B. Yegnanarayana, S. R. Mahadeva Prasanna, and K. Sreenivasa Rao, “Speech enhancement using excitation source information,” in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, pp. 541–544, May, 2002.

[2] N. D. Gaubitch and D. B. W. P A. Naylor, “On the use of linear prediction for dereverberation of speech,” in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, pp. 99–102, 2003.

[3] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations.* The Hague, The Netherlands: Mounton, 1970.

[4] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* Upper Saddle River, NJ: Prentice Hall, 2002.

[5] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[6] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Upper Saddle River, NJ: Prentice Hall, 1993.

[7] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, “Hmm-based speech synthesiser using the lf-model of the glottal source,” in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, pp. 4704–4707, 2011.

[8] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, M. Vainio, and P. Alku, “Hmm-based speech synthesis utilizing glottal inverse filtering,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 153–165, 2011.

[9] P. Hedelin, “High quality glottal lpc-vocoding,” in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, vol. 11, pp. 465–468, 1986.

[10] K. Cummings and M. Clements, “Analysis of the glottal excitation of emotionally styled and stressed speech,” *J. Acoust. Soc. Am.*, vol. 98, no. 1, pp. 88–98, 1995.

[11] T. Waaramaa, A. M. Laukkanen, M. Airas, and P. Alku, “Perception of emotional valences from vowel segments of continuous speech,” *Journal of Voice*, vol. 24, no. 1, pp. 30–38, 2010.

[12] Y. Koike and J. Markel, “Application of inverse filtering for detecting laryngeal pathol-

ogy," *The Annals of Otology, Rhinology, and Laryngology*, vol. 84, pp. 117–124, 1975.

[13] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–586, 1999.

[14] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 309–319, 1979.

[15] V. C. Raykar, B. Yegnanarayana, S. R. Mahadeva Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, no. 5, pp. 751–761, 2005.

[16] H. Hurley and R. Scott, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[17] A. H. Gray, Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 380–391, 1976.

[18] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing," in *Proc. Interspeech Conf.*, pp. 1353–1356, 2008.

[19] M. Lankarrany, W. P. Zhu, and M. N. S. Swamy, "Accurate estimation of the glottal flow derivative using iteratively reweighted 1-norm minimization," in *Proc. IEEE 9th Intl. Conf. in Circuits and Systems. (NEWCAS)*, pp. 33–36, 2011.

[20] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communication in Statistics-Theory and Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[21] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 1, 1993.

[22] P. Alku, E. Vikman, and U. K. Laine, "Analysis of glottal waveform in different phonation types using the new iaif-method," in *Proc. European Signal Processing Conf. (EUSIPCO)*, vol. 4, pp. 362–365, 1991.

[23] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[24] D. Y. Wong, J. D. Markel, and A. H. Gray., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 350–355, 1979.

[25] D. Andrews, P. Bickel, F. Hampel, P. Huber, and W. Rogers, "Robust estimates of location: Survey and advances," 1972.

[26] A. I. Koutrouvelis, R. Heusdens, and N. D. Gaubitch, "Comparison of linear prediction methods in terms of sparsity, stability and robustness to reverberation," in *Proc. 35th WIC Symposium on Information Theory in the Benelux*, pp. 156–163, 2014.

[27] D. Giacobello, M. G. Christensen, N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1644–1657, 2012.

[28] G. Kafentzis, "On the inverse filtering of speech," master thesis, University of Crete, Dept. of Computer Science, 2010.

[29] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-*

QPSR, vol. 26, no. 4, pp. 1–13, 1985.

[30] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 583–590, 1971.

[31] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds.," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1249–1260, 1995.

[32] J. L. Kelly and C. Lochbaum, "Speech synthesis," in *Proceedings of the fourth international congress on acoustics*, vol. Paper G42, pp. 1–4, 1962.

[33] B. H. Story, *Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract.* Doctoral dissertation, University of Iowa, 1995.

[34] B. H. Story and I. R. Titze, "Vocal tract area functions from magnetic resonance imaging.," *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 537–554, 1996.

[35] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Upper Saddle River, NJ: Prentice Hall, 1978.

[36] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals.* IEEE Press, 2000.

[37] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Communication*, vol. 1, no. 3-4, pp. 167–184, 1982.

[38] S. Maeda, "Vocal tract acoustics demonstrator." http://www.phon.ucl.ac.uk/resource/vtdemo/.

[39] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2B, pp. 637–655, 1971.

[40] J. Flanagan, *Speech Analysis, Synthesis, and Perception.* New York: Springer, 1972.

[41] M. Rothenberg, "An interactive model for the voice source," *STL-QPSR*, vol. 22, no. 4, pp. 1–17, 1981.

[42] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory.," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, 2008.

[43] G. F. M. Bavegard, "Notes on glottal source interaction ripple," *STL-QPSR*, vol. 35, no. 4, pp. 63–78, 1994.

[44] P. Alku, "Glottal inverse filtering analysis of human voice production - a review of estimation and parameterization methods of the glottal excitation and their applications.," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[45] W. R. Gardner and B. D. Rao, "Noncausal all-pole modeling of voiced speech.," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 1–10, 1997.

[46] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing 2nd Edition.* Upper Saddle River, NJ: Prentice Hall, 1998.

[47] G. Fant, "The lf-model revisited. transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[48] R. Veldhuis, "A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 566–571, 1998.

[49] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, vol. 11, pp. 1605–1608, 1986.

[50] T. V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics," *STL-QPSR*, vol. 25, no. 2-3, pp. 1–24, 1984.

[51] B. Doval and C. d'Alessandro, "The spectrum of glottal flow models," *Acta Acoustica United with Acustica*, vol. 92, pp. 1026–1046, 2006.

[52] I. R. Titze, "Theory of glottal airflow and source-filter interaction in speaking and singing," *Acta Acoustica United with Acustica*, vol. 90, pp. 641–648, 2004.

[53] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: an analytic study," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, pp. 1295 – 1298, 1997.

[54] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice," *J. Acoust. Soc. Am.*, vol. 84, no. 2, pp. 511–529, 1988.

[55] B. Bozkurt and T. Dutoit, "Mixed-phase speech modeling and formant estimation, using differential phase spectrums," in *VOQUAL*, pp. 21–24, 2003.

[56] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, pp. 855–866, 2011.

[57] A. Kounoudes, P. A. Naylor, and M. Brookes, "The dypsa algorithm for estimation of glottal closure instants in voiced speech.," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, vol. 1.

[58] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech Conf.*, 2009.

[59] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 994–1006, 2012.

[60] I. R. Titze and B. H. Story, "Rules for controlling low-dimensional vocal fold models with muscle activities.," *J. Acoust. Soc. Am.*, vol. 112, no. 3, pp. 1064–1076, 2002.

[61] B. Story, "Letalker 1.2." `http://sal.shs.arizona.edu/~bstory/LeTalkerMain.html`, 2013.

[62] J. J. A. P. A. Sluijter, V J. van Heuven, "Spectral balance as a cue in the perception of linguistic stress," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 503–513, 1979.

[63] G. Fant, "Glottal source and excitation analysis," *STL-QPSR*, vol. 20, no. 1, pp. 85–107, 1979.

[64] H. Wakita, "Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods," *SCRL Monograph*, no. 9, 1972.

[65] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: Wiley, 1996.

[66] H. Morikawa and H. Fujisaki, "Adaptive analysis of speech based on a pole-zero representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 1, pp. 77–88, 1982.

[67] J. L. Shanks, "Recursion filters for digital processing," *Geophysics*, vol. 32, pp. 33–51, 1967.

[68] K. Steiglitz, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 229–234, 1977.

[69] E. Denoël and J. P. Solvay, "Linear prediction of speech with a least absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 1397–1403, 1985.

[70] V. Khanagha and K. Daoudi, "An efficient solution to sparse linear prediction analysis of speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 3, 2013.

[71] J. E. Markel and A. H. Gray, *Linear prediction of speech.* New York: Springer, 1976.

[72] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Joint estimation of short-term and long-term predictors in speech coders," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, pp. 4109–4112, 2009.

[73] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Joint estimation of short-term and long-term predictors in speech coders," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, pp. 4650–4653, 2010.

[74] J. D. Markel, "Digital inverse filtering-a new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20, no. 2, pp. 129–137, 1972.

[75] A. H. Gray, JR. and J. D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 3, pp. 207–216, 1974.

[76] T. D. Drugman, "Maximum phase modeling for sparse linear prediction of speech," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 185–189, 2014.

[77] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, pp. 4650–4653, 2010.

[78] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3289–3305, 2009.

[79] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Am.*, vol. 56, pp. 1625–1629, 1974.

[80] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd edition.* New York: Springer, 2009.

[81] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, 2008.

[82] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 103–106, 2010.

[83] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Jap.*, vol. 53-A, pp. 36–43, 1970.

[84] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[85] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 492–501, 2006.

[86] P. Alku, M. Airas, and B. Story, "Evaluation of an inverse filtering technique using physical modeling of voice production," in *Proc. Interspeech Conf.*, p. 497âĂŞ500, 2004.

[87] G. Lindsey, A. Breen, and S. Nevard, "Spar's archivable actual-word databases," tech. rep., Univ. College London, London, UK, 1987.

[88] E. Candes and J. Romberg, "L1-magic: Recovery of sparse signals via convex programming," tech. rep., California Inst. Technol., Pasadena, CA, 2005.

[89] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, no. 4, pp. 943–950, 1979.

[90] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Am.*, vol. 80, no. 5, no. 5, pp. 1527–1529, 1986.

[91] T. D. T. Drugman, N. d'Alessandro, A. Moinet, and T. Dutoit, "Voice source parameters estimation by fitting the glottal formant and the inverse filtering open phase," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2008.

Glossary

List of Acronyms

SFM	source filter model
WGN	white Gaussian noise
SNR	signal to noise ratio
LTI	linear time invariant
MA	moving average
LP	linear prediction
LPC	linear prediction coefficients
GCI	glottal closure instant
GOI	glottal opening instant
LF	Liljencrants-Fant
ISD	Itakura-Saito distance
LSD	log spectral distortion distance
LBG	Linde-Buzo-Gray
LLSE	linear least squares estimator
MLE	maximum likelihood estimator
ZZT	zero Z transforms
LAD	least absolute deviations
DFT	discrete Fourier transform
IAIF	iterative adaptive inverse filtering
SGPC	sparse glottal pulse cancellation
FIR	finite impulse response
MRI	magnetic resonance imaging
ROC	region of convergence
FFT	fast Fourier transform
IRLSM	iteratively reweighted least squares method
AM	autocorrelation method

CM	covariance method
CPCM	closed-phase covariance method
WLPM	weighted linear prediction method
SLPM	sparse linear prediction method
IRSLPM	iteratively reweighted sparse linear prediction method
WELPM	weighted epoch linear prediction method
IAIFGC	iterative adaptive inverse filtering glottal cancellation

Index